ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Mar. 1999, p. 439–446
0066-4804/99/$04.00+0

Vol. 43, No. 3

MINIREVIEW

Genomics and Antimicrobial Drug Discovery

DONALD T. MOIR,[1] KAREN J. SHAW,[2] ROBERTA S. HARE,[2] AND GERALD F. VOVIS[1]*

Pathogen Genetics Department, Genome Therapeutics Corporation, Waltham, Massachusetts 02453-8443,[1] and Chemotherapy and Molecular Genetics, Schering-Plough Research Institute, Kenilworth, New Jersey 07033-0539[2]

INTRODUCTION

The increasing frequency of nosocomial infections due to methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE) and the fear that high-level vancomycin resistance will eventually spread to staphylococci underscore the need for vigilance in the continuing war against pathogenic microbes (18, 39). Current widely used antibiotics are targeted at a surprisingly small number of vital cellular functions: cell wall, DNA, RNA, and protein biosynthesis (Table 1), and instances of resistance to these antibiotics are widespread and well documented (48). Thus, there is little doubt that new antibiotics are needed to combat the growing problem of antibiotic-resistant bacteria, and targeting of new pathways will likely play an important role in discovery of these new antibiotics. In fact, a number of crucial cellular pathways, such as secretion, cell division, and many metabolic functions, remain untargeted today. In the last 3 years, high-throughput automated random genomic DNA sequencing together with robust fragment assembly tools has delivered a wealth of genomic sequence information to assist in the search for new targets. In many cases, entire biochemical pathways can be reconstructed and compared in different pathogens. The purpose of this minireview is to indicate where this information can be found, to outline some of the ways in which it can be used, and to describe new tools to take advantage of genomic sequence information in the drug discovery process.

Each potential new antibiotic must meet a number of criteria before it is approved for use, and the choice of an appropriate target is the first step in this process. It is helpful to review the utility of genomic information with regard to some of the key criteria which antimicrobial targets must meet. In general, (i) a target should provide adequate selectivity and spectrum, yielding a drug which is specific or highly selective against the microbe with respect to the human host but also active against the desired spectrum of pathogens; (ii) a target should be essential for growth or viability of the pathogen, at least essential under conditions of infection; and (iii) something about the function of the target should be known so that assays and high-throughput screens can be built. Identification of potential new targets can proceed from any one of these criteria, but ultimately all must be met by a successful target. For example, a variety of methods may be used to find genes which are essential for the survival of an organism under defined conditions or which are necessary for infectivity in an animal model. Comparative genomics may be used to identify potential targets which are shared across multiple microbial species. Several tools, primarily sequence similarity based, may be used to predict the function of most genes so that specific pathways can be targeted. As discussed below, genomic sequence information provides assistance in all of these areas: selectivity, spectrum, functionality, and essentiality (Fig. 1).

CURRENT RESOURCES FOR GENOMIC SEQUENCE AND FUNCTIONALITY INFORMATION

Numerous databases are now available which contain both sequence and functionality information. Most of these are accessible over the Internet through convenient Web browser interfaces. Many also permit downloading of sequence information for use on local servers. Sequence databases now contain the nucleotide and predicted amino acid sequences of virtually every gene in the model microbes *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae* as well as in a variety of other bacteria (Table 2; a version of this table is updated regularly by The Institute for Genomic Research [TIGR] on their Web site: http://www.tigr.org/tdb/mdb/mdb.html). These databases are the result of extensive analysis of the genomic sequences of those organisms. Open reading frames have been analyzed by sequence comparison and by codon usage to identify those which are most likely to represent transcribed genes. Putative functions have been assigned to slightly more than half of the genes in the model organisms based on sequence comparisons to genes of known function in other organisms, shared sequence motifs, or clustering of sequences into related families. Databases such as EcoCyc, KEGG, and WIT present these data in an organized and useful manner (see Table 3).

Recently, some commercial databases have also become available for nonexclusive use by commercial subscribers. These databases generally also provide sequence information not available in public databases and comparative software and analysis tools for convenient analysis of the data. For example, the results of prerun sequence similarity searches may be stored to provide rapid answers to complex comparative genomic queries by a subscriber. Finally, several Web-accessible sites offer useful tools for sequence analysis via sequence similarity searches, motif searches, and structural comparisons. Examples of relevant Internet sites providing databases of sequence and functionality information and research tools are described in Table 3.

The next advance in microbial genomics will be the availability of the complete genomic sequence from multiple strains of a single bacterial pathogen. The discovery of genes conserved in multiple pathogenic strains or the recognition of genes found only in the most virulent strains are examples of the power such genomic comparisons will provide. Sequence for a second strain of *Helicobacter pylori* has appeared and

* Corresponding author. Mailing address: Genome Therapeutics Corporation, 100 Beaver St., Waltham, MA 02453-8443. Phone: (781) 398-2313. Fax: (781) 398-2476. E-mail: jerry.vovis@genomecorp.com.

TABLE 1. Gene targets of widely used antibiotics

Target category and gene product	Antibiotic class
Protein synthesis	
30S ribosomal subunit	Aminoglycosides, tetracyclines
50S ribosomal subunit	Macrolides, chloramphenicol
$tRNA^{Ile}$ synthetase	Mupirocin
Elongation factor G	Fusidic acid
Nucleic acid synthesis	
DNA gyrase A subunit; topoisomerase IV	Quinolones
DNA gyrase B subunit	Novobiocin
RNA polymerase beta subunit	Rifampin
DNA	Metronidazole
Cell wall peptidoglycan synthesis	
Transpeptidases	Beta-lactams
D-Ala-D-Ala ligase substrate	Glycopeptides
Antimetabolites	
Dihydrofolate reductase	Trimethoprim
Dihydropteroate synthesis	Sulfonamides
Fatty acid synthesis	Isoniazid

sequence for a second strain of *Mycobacterium tuberculosis* will appear soon (Table 2).

COMPARATIVE GENOMICS TO ASSESS THE SPECTRUM AND SELECTIVITY OF A TARGET

One powerful use of genomic sequence information is to compare all of the identified genes in different bacterial pathogens to determine which genes are, or are not, shared by various species. Indeed, Tatusov et al. (50) have suggested that gene families conserved among bacteria but missing from eukaryotes comprise a pool of potential targets for broad-spectrum antibiotic development. An early step in this direction was taken by Mushegian and Koonin (36), who identified 256 genes shared by the two completely sequenced bacterial genomes at that time, those of *Haemophilus influenzae* and *Mycoplasma genitalium*. On the other hand, genes which are apparently unique to a species such as *H. pylori* might be ideal for targeting that species with a narrow-spectrum antibiotic. As the number of sequenced bacterial and fungal genomes grows, so does the ability to find genes common to most microbial pathogens or truly unique to a particular species. For example, Arigoni et al. (6) identified 26 genes in *E. coli*, most of which were conserved in the *B. subtilis*, *M. genitalium*, *H. influenzae*, *H. pylori*, *Streptococcus pneumoniae*, and *Borrelia burgdorferi* genomes. They reasoned that this list of genes, which had no predictable function, contained novel targets for broad-spectrum antibiotic development. These analyses can be extended by including sequence comparisons to eukaryotic genomes as a means to examine potential selectivity of a target (50). For example, Arigoni et al. (6) reported that 15 of 26 proteins broadly conserved across bacterial species also exhibited significant sequence similarity to proteins in *S. cerevisiae* and, therefore, represented targets which, in an assay, might identify compounds that also have human toxicity. While these targets could simply be avoided, it should be noted that the targets of the majority of marketed antimicrobial agents show some conservation with mammalian proteins.

As in all sequence comparisons, the search parameters and the quality of the input data, e.g., partial human or mammalian sequence information, are critical. Relevant issues which must be addressed include questions such as the following. What degree of sequence similarity to another bacterial genome indicates a shared gene? What degree of sequence similarity to a mammalian gene warns of a possible toxicity problem? Since sequence similarity-searching algorithms allow nearly complete flexibility in the choice of these parameters, some known examples are necessary to calibrate the method. Mushegian and Koonin (36) used a BLASTP score of 90 as the cutoff for defining a biologically relevant relationship between two protein sequences. The appropriate cutoff score for exclusion of genes with apparent mammalian homologs may be more gene specific. Some examples reveal a general trend. Trimethoprim is a highly selective inhibitor of bacterial dihydrofolate reductases (DHFR) despite the fact that the human and *E. coli* DHFR gene products share 28% amino acid identity over the length of the two proteins (40). Similarly, the quinolones are highly selective against bacterial gyrases despite the fact that the C-terminal domain of human topoisomerase II shares 20% amino acid identity with *E. coli* gyrase A (25). Fluconazoles are highly selective for fungal lanosterol 14-α demethylases, even though the human and yeast gene products share 37% amino acid identity over their full length (5). These sequence identity percentages translate into BLASTP scores of 132, 125, and 301, respectively, in a search of a large nonredundant protein database comprised of sequences from GenBank, SwissProt, and PIR. Therefore, exclusion of genes having apparent mammalian homologs with scores >150 would likely be suitable for a search of bacterial targets, but the score cutoff would have to be raised to allow identification of the broadest set of antifungal target genes.

IDENTIFICATION OF ESSENTIAL TARGETS EXPERIMENTALLY

Genomic sequence information is not required for discovering essential genes, but such information does facilitate the process. Genes which are essential to pathogenesis and prevent

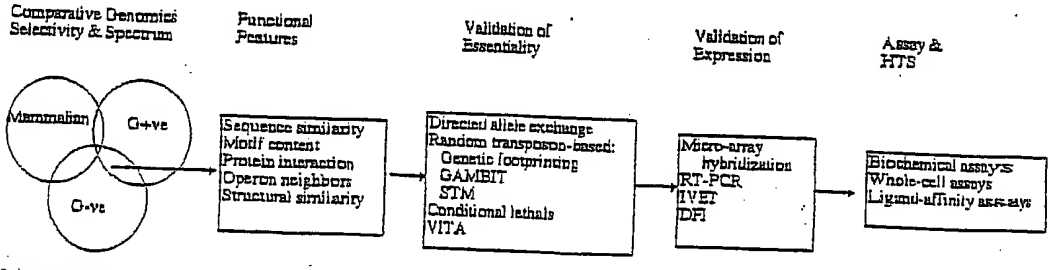


FIG. 1. Schematic view of genomic tools applied to antimicrobial-drug discovery. See the text for details. G+ve and G-ve, gram positive and gram negative, respectively.

TABLE 2. Sequenced microbial genomes

Internet resource	Genome	Strain(s)	Size (Mb)	Institution(s)	Reference
www.tigr.org/tdb/mdb/hidb/hidb.html	*Haemophilus influenzae* RD	KW20	1.83	TIGR	13
www.tigr.org/tdb/mdb/mgdb/mgdb.html	*Mycoplasma genitalium*	G-37	0.58	TIGR	15
www.tigr.org/tdb/mdb/mjdb/mjdb.html	*Methanococcus jannaschii*	DSM 2661	1.66	TIGR	8
www.kazusa.or.jp/cyano/cyano.html	*Synechocystis* sp.	PCC 6803	3.57	Kazusa DNA Research Institute	27
www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html	*Mycoplasma pneumoniae*	M129	0.81	University of Heidelberg	23
speedy.mips.biochem.mpg.de/mips/yeast/yeast_genome.html or genome-www.stanford.edu/Saccharomyces	*Saccharomyces cerevisiae*	S288C	13	European and North American Consortium	17
www.tigr.org/tdb/mdb/hpdb/hpdb.html	*Helicobacter pylori*	26695	1.66	TIGR	51
www.genetics.wisc.edu/	*Escherichia coli*	K-12	4.6	University of Wisconsin	7
www.genomecorp.com/gene/sequences/methanobacter/abstract.html	*Methanobacterium thermoautotrophicum*	delta H	1.75	Genome Therapeutics and Ohio State University	43
www.pasteur.fr/Bio/SubtiList.html	*Bacillus subtilis*	168	4.2	International Consortium	31
www.tigr.org/tdb/mdb/afdb/afdb.html	*Archaeoglobus fulgidus*	VC-16, DSM4304	2.18	TIGR	29
www.tigr.org/tdb/mdb/bbdb/bbdb.html	*Borrelia burgdorferi*	B31	1.44	TIGR	14
www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=133	*Aquifex aeolicus*	VF5	1.55	Diversa	10
www.bio.nite.go.jp/ot3db_index.html	*Pyrococcus horikoshii*	OT3	1.80	National Institute of Technology and Evaluation	28
www.sanger.ac.uk/Projects/M_tuberculosis/	*Mycobacterium tuberculosis*	H37Rv	4.40	Sanger Centre	9
www.tigr.org/tdb/mdb/tpdb/tpdb.html	*Treponema pallidum*	Nichols	1.14	TIGR and University of Texas	16
chlamydia-www.berkeley.edu:4231/	*Chlamydia trachomatis*	Serovar D (D/UW3/Cx)	1.05	University of California at Berkeley and Stanford University	46
evolution.bmn.uu.se/~siv/gnomics/Rickettsia.html	*Rickettsia prowazekii*	Madrid E	1.11	University of Uppsala	4
www.genomecorp.com/hpylori or www.astra-boston.com/hpylori	*Helicobacter pylori*	J99	1.64	Genome Therapeutics and Astra AB	3
www.tigr.org/cig-bin/BlastSearch/blast.cgi?organism=m_tuberculosis	*Mycobacterium tuberculosis*	CSU#93	4.40	TIGR	Unpublished

colony formation in a conditional-lethal manner are potential targets for new antimicrobials. This assumes that a small organic molecule which inhibits the activity of an essential gene product would either kill or inhibit the growth of the bacterium which requires that functional protein. Such conditional lethal genes can be discovered through classical mutagenesis techniques. Availability of the sequence of the genome means that the full sequence of each mutated gene, and frequently its cellular role as well, can be gleaned from a short sequence read on a complementing plasmid insert. This additional information accelerates the processing of a mutational study enormously. Depending on the availability of genetic tools for the microbial species in question, a variety of molecular genetic methods can be used to discover essential genes. For example, in *E. coli*, genes can be placed under control of a regulated promoter by use of an appropriately constructed transposon system (11), or genes can be mutated to a conditional-lethal form. In principle, such conditional mutants can be used in whole-cell screens under moderately suppressing conditions in which the cells may be hypersensitive to drug-like compounds which act against that gene product (see below).

It seems reasonable to assume that most genes which are essential to the cell for growth or viability on laboratory media will also be required for growth or viability in an infected host. Experimentally, media can be varied in order to identify genes which are essential under the widest range of growth conditions and particularly in rich media which may simulate conditions in necrotic tissue of an animal host. Cells carrying auxotrophic mutations may find sufficient nutritional supplement in the host tissues to permit growth or at least survival. Such genes might be poor targets for new antimicrobials unless experiments establish that the particular nutrient is in short supply in the host or that cells are incapable of transporting the nutrient efficiently. In order to establish that a gene target is essential in an infection, a transposon-based gene tagging method called "signature-tagged mutagenesis" (STM) has been used to identify genes which are essential in an animal model (22, 35). However, since cells carrying the disrupted tagged genes must be grown in the laboratory prior to introduction into the animal, the method may be biased against genes which are essential for growth both on laboratory media and in an animal model. Indeed, many of the genes identified by STM appear to encode virulence factors which affect the ability of the pathogen to colonize or damage host tissue rather than the viability of the pathogen. New drugs which intervene in these processes could prove highly selective, and resistance to such drugs might be rare since loss or mutation of the virulence factor would also likely reduce virulence. However, other resistance mechanisms, such as drug modification and efflux pumps, could be problematic. In addition, the absence of a convenient in vitro assay for such drugs would hamper the development, testing, and approval processes. It remains unclear how many important antimicrobial targets would be missed by using as targets for drug discovery only those genes which are essential for growth or viability on laboratory culture media.

A related, important feature of a suitable antimicrobial gene target is its expression pattern in the infection. The absolute level of expression may be less important than information about whether it is expressed at all. A highly expressed, abundant gene product should be no more difficult to inhibit than a low-abundance gene product since an inhibitor with suitably high affinity will be effective in either case unless it is poorly taken up by pathogens. However, if a gene is not expressed at all in an established infection of an animal host, then it will be of no interest as a potential target. A gene already established as being essential for growth or viability in the laboratory by genetic methods obviously must be expressed under these conditions because its failure to be expressed as an active product causes the pathogen to die. Knowledge that such an essential

gene is also expressed in an animal model would suggest that it is essential in an infection as well. Two types of methods offer information about gene expression. First, for genes whose sequence is known, reverse transcriptase PCR (RT-PCR) may be used to detect transcripts in cells grown on agar media or in animal infection models (47). Alternatively, for organisms which have been sequenced in their entirety, a whole-genome view of gene expression may be obtained by gridding clones, PCR products, or synthetic oligonucleotides representing every gene onto a solid support. Total RNA may be isolated from cells grown under conditions of interest, labeled, and hybridized to the array (12). While thorough, this type of method suffers from some problems: (i) appropriate controls must be run to eliminate the possibility of bacterial DNA contamination in the RNA preparation, (ii) probes are difficult to prepare because bacterial mRNA is notoriously unstable, and (iii) the whole-genomic scale of the experiments makes the arrayed membranes difficult and expensive to prepare and read. A genetic promoter trap method termed "in vivo expression technology" or IVET may be more feasible for most laboratories (21, 33). In this approach, which has been developed for use in *Salmonella typhimurium* grown intraperitoneally in BALB/c mice or in cultured macrophages, random DNA fragments are cloned upstream from a gene whose expression is required for growth in an animal host. Cells, which multiply in vivo, are recovered and cloned. The sequences of fragments serving as functional promoters in vivo are then determined. A second, related promoter trap method termed "differential fluorescence induction" (DFI) has been described recently (53). The distinguishing features of this approach are that (i) the gene used for selection encodes a modified green fluorescent protein and (ii) the selection is accomplished with a fluorescence-activated cell sorter. If such methods can be extended to other bacterial species and animal hosts, they will be extremely useful for assessing random genomic fragments or specific genes of interest for expression in vivo.

IDENTIFICATION OF ESSENTIAL TARGETS USING DATABASES

Potential gene targets selected from databases can be validated by examining the effect of a gene knockout on cell growth or viability. Recombination is almost exclusively between homologous regions in bacterial genomes, and many common pathogens as well as model bacteria are transformable. Exchange between the chromosomal wild-type allele and a version engineered to carry a deletion and/or an insertion of a drug resistance cassette is generally efficient enough to be practical in the laboratory. Interpreting the results of such an experiment, however, may be difficult for two reasons. First, the frequent occurrence of polycistronic messages in bacteria means that disruption of a gene may have a deleterious effect on expression of a distal neighboring gene, a so-called "polar" effect. In that case, the inviability caused by a gene knockout could be due to loss of expression of a gene other than the one disrupted. Precautions can be taken to reduce these effects by, for example, including a moderate-strength outward reading promoter in the disrupted version of the allele so as to permit expression of the downstream gene(s). Second, the method works better as an exclusionary tool than as an inclusionary one. While success in generating a cell carrying a disrupted allele indicates that the gene is not essential for growth or viability of the cell, failure to generate such an altered cell could be due to any one of multiple causes including polar effects or inefficient recombination in a particular genetic interval.

TABLE 3. Additional Internet resources

Database or organization	Internet address
Sequence databases	
NCBI	http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
DDBJ	http://www.ddbj.nig.ac.jp/htmls_test/Welcome-e.html
EBI/EMBL	http://www.ebi.ac.uk/ebi_home.html
GSDB	http://www.ncgr.org/gsdb/Index_gsdb.html
SwissProt (Geneva)	http://expasy.hcuge.ch/www/expasy-top.html
Candida	http://alces.med.umn.edu/Candida.html
MIPs	http://www.mips.biochem.mpg.de/
RDP	http://rdp.life.uiuc.edu/
SGD	http://genome-www.stanford.edu/
Metabolic databases	
KEGG	http://www.genome.ad.jp/kegg/
Ecocyc	http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html
WIT	http://www.cme.msu.edu/WIT/
Sequencing groups	
Berkeley	http://chlamydia-www.berkeley.edu:4231/
Genome Therapeutics	http://www.genomecorp.com/home.htl
Sanger	http://www.sanger.ac.uk/Projects/
Stanford	http://sequence-www.stanford.edu/group/malaria/index.html
TIGR	http://www.tigr.org/tdb/mdb/mdb.html
University of Oklahoma	http://dna1.chem.uoknor.edu/index.html
University of Queensland	http://www.cmcb.uq.edu.au/aeruginosa/
University of Washington	http://chimera.biotech.washington.edu/uwgc/
Washington University	http://genome.wustl.edu/gsc/bacterial/salmonella.html
Tools and resources	
Biomolecular Research Tools	http://www.public.iastate.edu/~pedro/rt_1.html
COGs	http://www.ncbi.nlm.nih.gov/COG/
NCGR	http://www.ncgr.org/microbe/Index_bms.html
MAGPIE	http://www.mcs.anl.gov/home/gaasterl/genomes.html
Genobase	http://spector.dcrt.nih.gov:8004/
Micro Underground	http://www.isume.edu/campus/micr/mirror/public_html/index.html
ANMR	http://www.wdcm.riken.go.jp/
WHO	http://www.who.ch/Welcome.html
Pallen	http://www.qmw.ac.uk/~rh.bm001/textbook/chapter.html
CDC	http://www.cdc.gov/
University of Kansas	http://www.ikume.edu/research/fgsc/main.html
University of Georgia	http://fungus.genetics.uga.edu:5080/
Tripos	http://www.tripos.com/sites.html
Motif	http://dna.Stanford.EDU/identify/
Pedant	http://pedant.mips.biochem.mpg.de/
GenTHREADER	http://globin.bio.warwick.ac.uk/genome/genomic.html

One solution to this problem is to carry out allele exchange as a two-step process (20, 32). In *E. coli*, for example, the disrupted allele together with the vector carrying it can be integrated into the genome by means of a single crossover, a so-called "Campbell insertion." Recombination between homologous regions on the two copies of the allele now on the chromosome will eliminate the vector sequences and either copy of the allele. Which copy is eliminated depends upon which regions of homology were involved in the recombination. Failure to find cells retaining only the disrupted allele strongly suggests that such progeny are inviable. Success in finding cells retaining only the wild-type allele confirms that

recombination is efficient in this genetic interval. However, in many naturally competent bacterial species, such as *H. influenzae* and *S. pneumoniae*, double-crossover events are extremely efficient, and allele replacement occurs with little or no opportunity to isolate a single crossover intermediate (1). While this complicates evaluation of essential genes in these organisms, it provides a convenient method for disrupting genes under conditions in which they are not essential so that the resulting strains may be examined under a variety of other conditions (e.g., see below).

A new approach promises to accelerate the process of evaluating the essentiality of genes. Smith et al. (44, 45) have described a method for the yeast *S. cerevisiae* called "genetic footprinting" which makes use of a quasi-random transposable Ty element to generate a rich array of gene knockouts in a population of cells. Further transposition is shut off, and the population is then grown under a variety of conditions. DNA is prepared from cells in the various growth populations, and the DNA is queried by PCR amplification to determine if it will yield PCR products between a gene-specific primer and a transposon-specific primer. Failure to find such PCR products suggests that cells carrying transposons in that gene were inviable under the growth conditions employed. Fluorescent PCR products are viewed on standard sequencing gels by using automated fluorescence sequencing machines and a commercially available software package. An important control in this method is the existence of a gene-to-transposon PCR product in the so-called t_0 cell population prior to the shutdown of transposition. This assures the experimenter that this region is not simply a "cold" spot for transposition. The efficiency of this method derives from the use of random transposons to build all necessary gene knockouts rapidly, followed by automated PCR and analysis methods to interpret the results for any given gene of interest.

Recently, a modified version of this method, called "genomic analysis and mapping by in vitro transposition" (GAMBIT), has been applied successfully to two bacterial species (1). In this variation of genetic footprinting, the transposition mutagenesis was done on PCR-amplified genomic segments from *H. influenzae* or *S. pneumoniae* in vitro, and the mutations were introduced into these naturally competent host bacteria by transformation. While the method suffers from the absence of a true t_0, the focus on 10-kb DNA segments permits near-saturation mutagenesis with the *mariner* family transposon *Himar1*, which shows little or no insertion site specificity. These authors identified four essential conserved genes of unknown function from a total of 13 analyzed.

Currently, the main limitation to this method is a requirement for an efficiently transformable host bacterium so that mutations generated in vitro can be evaluated readily in vivo. Other limitations which apply to all genetic footprinting methods include the following: (i) essentiality of the function of a gene that is duplicated or has a functional paralog cannot be analyzed, since footprinting assesses the fitness of a single mutagenized gene; (ii) polarity effects, although not a problem for *S. cerevisiae*, may lead to misinterpretation of data obtained from bacteria; (iii) the correlation of footprinting data with gene knockout data has not been confirmed in any organism; and (iv) footprinting data are technically difficult to interpret for a variety of reasons, including the facts that some essential genes will tolerate insertions in the C-terminal coding region (e.g., *secA* [1]) and cells carrying insertions in some genes display an intermediate slow-growth phenotype (e.g., *ade2* [44]).

TOOLS FOR PREDICTING THE FUNCTION OF GENE PRODUCTS

Clearly, not all of the predicted functional assignments based on sequence similarities are reliable. In some cases, for example, the function of the closest-related protein has itself been predicted based on its sequence similarity to a gene product of known function. In other cases, the chain of relatedness to a protein of confirmed function may be even longer. About half of the genes in bacterial genomes either lack significant enough sequence similarity to permit functional assignment or have likely homologs whose function is unknown. In neither of these cases can a function be predicted for the gene product. Nevertheless, the results of sequence similarity searches are a useful starting point for further investigation. More sensitive sequence comparison searches may provide a putative function or functional feature such as the presence of a short protein sequence motif. For example, a search against a database of clusters of orthologous groups of genes (COGs [Table 3]) yielded over 100 additional functional predictions for genes in the *H. pylori* genome (50).

Tools other than sequence similarity have also been useful in a few cases for predicting function of a gene product. For example, a gene product, with no significant sequence relationship to a protein of known function but which is likely to be cotranscribed as part of a polycistronic message with other genes of known function, may play a role in the same pathway with the known gene products. In the *E. coli* genome, the hypothetical gene *yjaF* appears to be cotranscribed with the porphyrin biosynthetic gene *hemE*, and the hypothetical gene *yadM* appears to be in an operon with the outer-membrane usher protein HtrE, which is involved in transport and binding. It is reasonable to speculate that these genes of unknown function play roles in the same biochemical pathways as their neighboring "known" genes. Of course, experimental evidence would be required to confirm these hypotheses. Methods also exist for identifying likely structural similarity even in the absence of strong primary sequence similarity. As the databases of known structures grow, this will become a powerful approach for assigning likely functions to gene products. For example, the "GenTHREADER" web site (Table 3) presents analysis results from a fast fold recognition program on the predicted open reading frames from three bacterial genomes.

Laboratory methods can also be invoked to solve questions of unknown gene identities. An unknown gene may be used as the bait in a yeast two-hybrid interaction trap to identify genes whose protein products interact with the unknown protein. The identity of an interacting partner will frequently implicate the unknown in a particular cellular pathway (19). Finally, an unknown gene may be expressed as a tagged fusion, the protein purified by affinity column, and the product tested for categories of activities such as proteolysis, DNA cleavage or binding, ATP or GTP hydrolysis, and binding, to name a few. The probability of successfully identifying an activity of an unknown by the latter method is low, but this method may be warranted if sequence comparisons suggest the presence of a motif associated with an assayable function. An attractive alternative is to focus on assays which do not require knowledge of the cellular function of a gene product (see below).

THE FUTURE: DEALING WITH GENE TARGETS HAVING NO PREDICTABLE FUNCTIONAL FEATURES

The array of tools described so far, including comparative genomic methods for identifying potentially useful gene targets and allele exchange methods for validating the essentiality of

those genes, provides both gene targets whose cellular function can be predicted and gene targets for which little or no functional information is available. Targets in the first class may be used immediately to build biochemical assays and high-throughput screens to detect small organic molecules which inhibit the biochemical activity. Typically, the gene sequence is amplified by PCR from genomic DNA of a given bacterium, inserted into an expression vector, and expressed in *E. coli* sometimes with affinity tags to facilitate purification of the resulting protein product.

It is far less obvious how to proceed with gene targets lacking any functional information. This problem has attracted considerable attention in recent years because of the growing number of such targets known to be shared across many bacterial species (24), some of which are known to be essential in at least one species. As a general guide, about 40% of bacterial genes cannot be assigned a putative function at this time. If 10 to 15% of these genes are essential, then 4 to 6% of the genes in a typical bacterial genome (about 100 genes) represent potential antimicrobial targets which have never been used in screens. Three basic types of approaches seem feasible and have shared some initial success. First, cells expressing higher- or lower-than-normal levels of particular genes have in some cases been shown to be more resistant or more sensitive, respectively, than their wild-type parents to chemical compounds known to inhibit those gene products. For example, overexpression of the yeast *ALG7* gene results in cells more resistant than wild-type cells to tunicamycin (38), while reduced activity of the same gene product results in cells more sensitive to the drug (30). Similarly, increased expression of the *ERG11* gene in *Candida glabrata* results in higher levels of resistance to the azole family of drugs which target that enzyme (54). A gene of unknown function could be overexpressed in a host strain, and the resulting assay strain could be tested for increased resistance to a library of compounds. It is clear, however, that many gene targets when overexpressed do not lead to resistance to chemical compounds that are known to bind to the protein product (e.g., *gyrA* [52]). Furthermore, overexpression of proteins often leads to lethality or growth defects (e.g., *kasA* [34]). Alternatively, a gene could be underexpressed or crippled by a mutation so that cells might show increased sensitivity to a compound which inhibits the protein product. Scientists at Microcide Pharmaceuticals, Inc., have applied this approach on a large scale using temperature-sensitive mutants grown at intermediate temperatures in order to reduce the level of activity of the target gene product (39a). Of course, it is not clear what fraction of unknown gene products would provide the cell with increased drug resistance or sensitivity when over- or underexpressed in these ways.

The second approach to this problem of assaying gene products of unknown function is probably more generally applicable. Libraries of small molecules are screened for strong binding affinity to proteins of unknown function. This has been achieved with peptides in phage display libraries because binding can be readily detected by elution of bound phage from the protein tethered on a solid support. Proteins of unknown function can be produced easily as affinity fusion products for attachment to solid supports, and a variety of peptide phage display libraries are commercially available. Conformationally constrained disulfide-bonded peptides with affinities in the 100 μM to 100 nM range can be obtained by this approach (55). Of course, not all peptides detected by this approach will bind to sites which inhibit activity, but an elegant new method, called "validation in vivo of targets for anti-infectives" (VITA), has been devised to identify those peptides which inhibit essential cellular functions (49). Potential inhibitory peptides were expressed in a regulated manner within bacterial host cells which were grown either on agar medium or in an animal model of infection. Inhibition of cell growth or viability upon induction of peptide expression validated the peptide-protein interaction as useful for further drug development. While peptides are not ideal drug candidates, a wider array of techniques are applicable after a moderate binder has been obtained. The peptide may be used as a surrogate ligand in a competition assay to identify a small organic compound with higher affinity. Scintillation proximity assays (26) or fluorescence polarization assays (41) may be used in a high-throughput mode to identify compounds in chemical libraries which compete for binding with a labeled peptide. Alternatively, ligand binding assays may be configured to work directly on libraries of unlabeled chemical compounds. Shuker et al. (42) have described a nuclear magnetic resonance-based method capable of a throughput of 1,000 compounds per day. Mass spectrometric methods are also of interest as potentially rapid ways to detect bound ligands from chemical libraries. One concern about these approaches is that proteins may have multiple accessible binding sites, many of which have nothing to do with catalytic activity. It is not clear at this early stage how significant an issue multiple binding sites will be. However, it is worth noting that Shuker et al. (42) took advantage of a second binding site to increase the affinity of an inhibitor for the protein. Ultimately, of course, affinity ligands must be shown to inhibit cell growth, that is, to have antimicrobial activity. Some chemical engineering of the compound may be required to increase microbial uptake.

A third approach for assaying gene products of unknown function relies on the complex gene expression regulatory network found in many bacteria. Expression levels of genes in metabolic pathways are often regulated in response to the amounts of intermediates in the cell. For example, disruption of the general secretory pathway in *E. coli* by mutation results in dramatic up-regulation of *secA* gene expression (37). Alksne et al. (2) took advantage of this fact to build a strain of *E. coli* carrying a *secA-lacZ* fusion as a detectable reporter. Several synthetic compounds and natural products were identified by their ability to induce expression of the reporter. Many of these exhibited antimicrobial activity and reduced the secretion of *Staphylococcus aureus* toxin 1. Similarly, Mdluli et al. (34) have reported that sublethal concentrations of isoniazid lead to up-regulation of the *kasA* and *acpM* genes. This group has initiated a whole-cell, high-throughput screen of chemical compounds which induce expression of a luciferase reporter fused to a gene in this regulated pathway. Screens of this type, which take advantage of the bacterial gene regulatory network, are inherently less specific than the two other types described here. In addition, they suffer from the basic limitation of all whole-cell screens: compounds must be capable of entering the cell in order to be detected. However, these types of screens offer the potential advantage of identifying compounds which act at any of several points in a pathway.

CONCLUSIONS

The availability of genomic sequence information for all or nearly all of several different bacterial species provides important new advantages for target discovery. First, it permits use of a comparative genomic analysis to identify potential new targets shared across several bacterial species or particular to a single species. In this manner, it is possible to generate lists of genes which represent potential targets for broad-spectrum or highly focused narrow-spectrum antibiotics. Sequence comparisons can also provide some assurance against mammalian

toxicity if proteins of similar sequence do not exist in mammalian sequence databases. Second, sequence similarity provides some insights into putative functions for most gene products. Finally, availability of the entire sequence of the gene target of interest permits rapid construction of gene knockouts to validate the utility of the target and facile construction of expression plasmids for production of protein and development of assays. The fact that bacterial and fungal genes can be assessed rapidly for their relevance as potential antibiotic targets by determining the effect of knocking out the gene and the fact that their genomes are small enough to be sequenced in their entirety are compelling reasons that the field of genomics will likely find its first real utility in the development of new antimicrobials.

ACKNOWLEDGMENTS

We thank our colleagues at Genome Therapeutics Corporation and the Schering-Plough Research Institute for helpful discussions about genomic approaches to drug discovery. In particular, Skip Shimer, Brad Guild, and Lucy Ling were instrumental in the analysis of the approaches summarized here. We thank Douglas Smith of Genome Therapeutics Corporation for the compilation of Internet resources presented in Table 3.

REFERENCES

1. Akerley, B. J., E. J. Rubin, A. Camilli, D. J. Lampe, H. M. Robertson, and J. J. Mekalanos. 1998. Systematic identification of essential genes by *in vitro mariner* mutagenesis. Proc. Natl. Acad. Sci. USA 95:8927–8932.
2. Alksne, L. E., P. Burgio, P. Bradford, B. Feld, W. Hu, P. Labthavikul, M. McGlynn, P. J. Petersen, M. Tuckman, and S. Projan. 1998. Identification of inhibitors of bacterial secretion by using a SecA reporter system, p. 272. *In* Abstracts of the 38th Interscience Conference on Antimicrobial Agents and Chemotherapy. American Society for Microbiology, Washington, D.C.
3. Alm, R. A., L. L. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, B. Noonan, B. C. Guild, B. L. deJonge, G. Carmel, P. J. Tummino, A. Caruso, M. Uria-Nickelsen, D. M. Mills, C. Ives, R. Gibson, D. Merberg, S. D. Mills, Q. Jiang, D. E. Taylor, G. F. Vovis, and T. J. Trust. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176–180.
4. Andersson, S. G. E., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. M. Alsmark, R. M. Podowski, A. K. Naeslund, A.-S. Eriksson, H. H. Winkler, and C. G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140.
5. Aoki, Y., F. Yoshihara, M. Kondoh, Y. Nakamura, N. Nakayama, and M. Arisawa. 1993. Ro 09-1470 is a selective inhibitor of P-450 lanosterol C-14 demethylase of fungi. Antimicrob. Agents Chemother. 37:2662–2667.
6. Arigoni, F., F. Talabot, M. Peitsch, M. D. Edgerton, E. Meldrum, E. Allet, R. Fish, T. Jamotte, M.-L. Curchod, and H. Loferer. 1998. A genome-based approach for the identification of essential bacterial genes. Nat. Biotechnol. 16:851–856.
7. Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1462.
8. Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, and J. C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073.
9. Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544.
10. Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 393: 353–358.
11. de Lorenzo, V., L. Eltis, B. Kessler, and K. N. Timmis. 1993. Analysis of *Pseudomonas* gene products using *lacI*^{q}/*Ptrp-lac* plasmids and transposons that confer conditional phenotypes. Gene 123:17–24.
12. DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686.
13. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.
14. Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J.-F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580–586.
15. Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison III, and J. C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403.
16. Fraser, C. M., S. J. Norris, G. M. Weinstock, O. White, G. G. Sutton, R. Dodson, M. Gwinn, E. K. Hickey, R. Clayton, K. A. Ketchum, E. Sodergren, J. M. Hardham, M. P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J. K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M. D. Cotton, J. C. Venter, et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388.
17. Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. 1996. Life with 6000 genes. Science 274:546–567.
18. Gold, H. S., and R. C. Moellering. 1996. Antimicrobial-drug resistance. N. Engl. J. Med. 335:1445–1453.
19. Gyuris, J., E. Golemis, H. Chertkov, and R. Brent. 1993. Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2. Cell 75:791–803.
20. Hamilton, C. M., M. Aldea, B. K. Washburn, P. Babitzke, and S. R. Kushner. 1989. New method for generating deletions and gene replacements in *Escherichia coli*. J. Bacteriol. 171:4617–4622.
21. Helfhoff, D. M., C. P. Conner, P. C. Hanna, S. M. Julio, U. Hentschel, and M. J. Mahan. 1997. Bacterial infection as assessed by *in vivo* gene expression. Proc. Natl. Acad. Sci. USA 94:934–939.
22. Hensel, M., J. E. Shea, C. Gleason, M. D. Jones, E. Dalton, and D. W. Holden. 1995. Simultaneous identification of bacterial virulence genes by negative selection. Science 269:400–403.
23. Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24:4420–4449.
24. Hinton, J. C. D. 1997. The *Escherichia coli* genome sequence: the end of an era or the start of the FUN? Mol. Microbiol. 26:417–422.
25. Hoshino, K., K. Sato, T. Une, and Y. Osada. 1989. Inhibitory effects of quinolones on DNA gyrase of *Escherichia coli* and topoisomerase II of fetal calf thymus. Antimicrob. Agents Chemother. 33:1816–1818.
26. Jenh, C. H., M. Zhang, M. Wiekowski, J. C. Tan, X. D. Fan, V. Hegde, M. Patel, R. Bryant, S. K. Narula, P. J. Zavodny, and C. C. Chen. 1998. Development of a CD28 receptor binding-based screen and identification of a biologically active inhibitor. Anal. Biochem. 256:47–55.
27. Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3: 109–136.
28. Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, and H. Kikuchi. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5(Suppl.):147–155.
29. Klenk, H.-P., R. A. Clayton, J.-F. Tomb, O. White, K. E. Nelson, K. A.

Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weidman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fujii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370.

30. Kukuruzinska, M. A., and K. Lennon. 1995. Diminished activity of the first N-glycosylation enzyme, dolichol-P-dependent N-acetylglucosamine-1-P transferase (GPT), gives rise to mutant phenotypes in yeast. Biochim. Biophys. Acta 1247:51–59.
31. Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S.-K. Choi, J.-J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Düsterhöft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, et al. 1997. The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature 390:249–256.
32. Link, A. J., D. Phillips, and G. M. Church. 1997. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. J. Bacteriol. 179:6228–6237.
33. Mahan, M. J., J. W. Tobias, J. M. Slauch, P. C. Hanna, R. J. Collier, and J. J. Mekalanos. 1995. Antibiotic-based selection for bacterial genes that are specifically induced during infection of a host. Proc. Natl. Acad. Sci. USA 92:669–673.
34. Mdluli, K., R. A. Slayden, Y.-Q. Zhu, S. Ramaswamy, X. Pan, D. Mead, D. D. Crane, J. M. Musser, and C. E. Barry. 1998. Inhibition of a *Mycobacterium tuberculosis* β-ketoacyl ACP synthase by isoniazid. Science 280:1607–1610.
35. Mei, J. M., F. Nourbakhsh, C. W. Ford, and D. W. Holden. 1997. Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. Mol. Microbiol. 26:399–407.
36. Mushegian, A. R., and E. V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA 93:10268–10273.
37. Riggs, P. D., A. L. Derman, and J. Beckwith. 1988. A mutation affecting the regulation of a secA-lacZ fusion defines a new sec gene. Genetics 118:571–579.
38. Ring, J. 1991. Gene overexpression in studies of *Saccharomyces cerevisiae*. Methods Enzymol. 194:235–251.
39. Salyers, A. A., and C. F. Amabile-Cuevas. 1997. Why are antibiotic resistance genes so resistant to elimination? Antimicrob. Agents Chemother. 41:2321–2325.
39a. Schmid, M. Personal communication.
40. Schweitzer, B. I., A. P. Dicker, and J. R. Bertino. 1990. Dihydrofolate reductase as a therapeutic target. FASEB J. 4:2441–2452.
41. Seethala, R., and R. Menzel. 1997. A homogeneous, fluorescence polarization assay for src-family tyrosine kinases. Anal. Biochem. 253:210–218.
42. Shuker, S. B., P. J. Hajduk, R. P. Meadows, and S. W. Fesik. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. Science 274:1531–1534.
43. Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, H. Safer, D. Patwell, S. Prabhakar, S. McDougall, G. Shimer, A. Goyal, S. Pietrokovski, G. M. Church, C. J. Daniels, J.-I. Mao, P. Rice, J. Noalling, and J. N. Reeve. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. 179:7135–7155.
44. Smith, V., D. Botstein, and P. O. Brown. 1995. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc. Natl. Acad. Sci. USA 92:6479–6483.
45. Smith, V., K. N. Chou, D. Lashkari, D. Botstein, and P. O. Brown. 1996. Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science 274:2069–2074.
46. Stephens, R. S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, E. V. Koonin, and R. W. Davis. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759.
47. Swartley, J. S., L.-J. Liu, Y. K. Miller, L. E. Martin, S. Edupuganti, and D. S. Stephens. 1998. Characterization of the gene cassette required for biosynthesis of the (α1→6)-linked *N*-acetyl-D-mannosamine-1-phosphate capsule of serogroup A *Neisseria meningitidis*. J. Bacteriol. 180:1533–1539.
48. Swartz, M. N. 1994. Hospital-acquired infections: diseases with increasingly limited therapies. Proc. Natl. Acad. Sci. USA 91:2420–2427.
49. Tao, J., T. Li, G. Connelly, X. Shen, J. Silverman, F. Houman, P. Wendler, and F. P. Tally. 1998. VITA: validation in vivo of targets and assays for anti-infectives, p. 274. *In* Abstracts of the 38th Interscience Conference on Antimicrobial Agents and Chemotherapy. American Society for Microbiology, Washington, D.C.
50. Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science 278:631–637.
51. Tomb, J.-F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, M. D. Cotton, J. M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–542.
52. Truong, Q. C., J. C. Nguyen Van, D. Shlaes, L. Gutmann, and N. J. Moreau. 1997. A novel, double mutation in DNA gyrase A of Escherichia coli conferring resistance to quinolone antibiotics. Antimicrob. Agents Chemother. 41:85–90.
53. Valdivia, R. H., and S. Falkow. 1997. Fluorescence-based isolation of bacterial genes expressed within host cells. Science 277:2007–2011.
54. Vandenbossche, H., P. Marichal, F. C. Odds, L. Lejeune, and M. C. Coene. 1992. Characterization of an azole-resistant *Candida glabrata* isolate. Antimicrob. Agents Chemother. 36:2602–2610.
55. Wrighton, N. C., F. X. Farrell, R. Chang, A. K. Kashyap, F. P. Barbone, L. S. Mulcahy, D. L. Johnson, R. W. Barrett, L. K. Jolliffe, and W. J. Dower. 1996. Small peptides as potent mimetics of the protein hormone erythropoietin. Science 273:458–463.

A Genomic Perspective on Protein Families

Roman L. Tatusov, Eugene V. Koonin,* David J. Lipman

In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

The release in 1995 of the complete genome sequence of the bacterium *Haemophilus influenzae* (*1*), followed within the next 1.5 years by four more bacterial genomes (*2*), one archaeal genome (*3*), and one genome of a unicellular eukaryote (*4*), marked the advent of a new age in biology. The hallmark of this era is that comparisons between complete genomes are becoming an indispensable component of our understanding of a variety of biological phenomena. The number of sequenced genomes is expected to grow exponentially for at least the next few years, and conceivably, their impact on biology will further increase (*5*).

Knowing the inventory of conserved genes responsible for housekeeping functions and understanding the differences in the genetic basis of these functions in different phylogenetic lineages is central to understanding life itself, at least at the level of a single cell. Complete sequences are indispensable for achieving this goal because they hold the only type of information that can be used to delineate the complete network of relationships between genes from different genomes. Furthermore, only with complete genome sequences is it possible to ascertain that a particular protein implicated in an essential function is not encoded in a given genome. Accordingly, an alternative protein for the respective function should be sought among the functionally unassigned gene products (*6*). With multiple genome sequences, it is possible to delineate protein families that are highly conserved in one domain of life but are missing in the others. Such information may be critically important: For example, the families that are conserved among bacteria but are missing in eukaryotes comprise the pool of potential targets for broad-spectrum antibiotics.

The knowledge of all of the gene sequences from multiple complete genomes redefines the problem of gene classification. It becomes feasible to replace the more or less arbitrary clustering of genes by similarity with a complete, consistent system in which the groups are likely to have evolved from a single ancestral gene. Such a natural classification of genes will provide a framework for evolutionary studies and for rapid, largely automatic functional annotation of newly sequenced genomes. This framework will evolve and improve with increasing coverage of the diversity of life forms with complete genome sequences. It is critical to have this system in place while the number of completed genomes is still small and each family can be explored individually. Here we describe a prototype of a natural system of gene families from complete genomes.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. Orthologs are genes in different species that evolved from a common ancestral gene by speciation; by contrast, paralogs are genes related by duplication within a genome (*7*). Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original one. Thus, identification of orthologs is critical for reliable prediction of gene functions in newly sequenced genomes. It is equally important for phylogenetic analysis because interpretable phylogenetic trees generally can be constructed only within sets of orthologs (*8*). A complete list of orthologs also is a prerequisite for any meaningful comparison of genome organization (*9*).

A naïve operational definition would simply maintain that for a given gene from one genome, the gene from another genome with the highest sequence similarity is the ortholog. Given the complete genome sequences, this straightforward approach often gives credible results, especially when the compared species are not too distant phylogenetically (*9*). At larger phylogenetic distances, however, the situation becomes more complicated. If gene duplications occurred in each of the given two clades subsequent to their divergence, only a many-to-many relationship will adequately describe orthologs, and accordingly, detection of the highest similarity will not result in the identification of the complete set of orthologs. In addition, when the best hit is not highly significant statistically, which is common in the case of phylogenetically distant relationships (*10*), it simply may be spurious. On the other hand, attempts to apply a restrictive similarity cutoff are likely to result in a number of orthologs being missed.

Given the existence of one-to-many and many-to-many orthologous relationships, we redefined the task of identifying orthologs as the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages. In other words, any two proteins from different lineages that belong to the same COG are orthologs. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

In order to delineate the COGs, all pairwise sequence comparisons among the 17,967 proteins encoded in the seven complete genomes were performed (*11*), and for each protein, the best hit (BeT) in each of the other genomes was detected. The identification of COGs was based on consistent patterns in the graph of BeTs. The simplest and most important of such patterns is a triangle, which typically consists of orthologs (Fig. 1A). Indeed, if a gene from one of the compared genomes has BeTs in two other genomes, it is highly unlikely that the respective genes are also BeTs for one another unless they are bona fide orthologs (*12*). The consistency between BeTs resulting in triangles does not depend on the absolute level of similarity between the compared proteins and thus allows the detection of orthologs among both slowly and quickly evolving genes. This approach is most likely to be informative when the

The authors are with the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

*To whom requests for reprints should be addressed. E-mail: koonin@ncbi.nlm.nih.gov

EXHIBIT 2

BeTs forming a triangle come from widely different lineages. Accordingly, only five major, phylogenetically distant clades were used as independent contributors to COGs: Gram-negative bacteria (*Escherichia coli* and *H. influenzae*), Gram-positive bacteria (*Mycoplasma genitalium* and *M. pneumoniae*), Cyanobacteria (*Synechocystis* sp.), Archaea (Euryarchaeota) (*Methanococcus jannaschii*), and Eukarya (Fungi) (*Saccharomyces cerevisiae*) (*13*).

The procedure used to derive COGs included finding all triangles formed by BeTs between the five major clades and merging those triangles that had a common side until no new ones could be joined. A triangle is an elementary, minimal COG (Fig. 1A). The groups produced by merging adjacent triangles include orthologs from different lineages and, in many cases, paralogs from the same lineage (Fig. 1, B and C). Because of the existence of paralogs, the BeTs that form the triangles are not necessarily symmetrical: For example, in the COG shown in Fig. 1C, the same *M. genitalium* protein, MG249, is the BeT for four paralogous σ subunits of *E. coli* RNA polymerase, but only for one of them, RpoD, is the relationship symmetrical.

Most of the clusters derived by the above procedure meet the definition of a COG, that is, all of the proteins from the different lineages in the same cluster are likely to be orthologs. There are, however, several reasons why, in certain cases, COGs may be lumped together. Proteins may contain two or more distinct regions, each of which belongs to a different conserved family; usually such proteins are loosely referred to as multidomain (*14*). Each of the clusters was inspected for the presence of multidomain proteins, individual domains were isolated (*15*), and a second iteration of the sequence comparison was performed with the resulting database of domains. Some of the COGs may include proteins from different lineages that are paralogs rather than orthologs, primarily because of differential gene loss in the major phylogenetic lineages. When one gene in a pair of paralogs is lost in one lineage but not in the others, two COGs that should have been distinct may be artificially joined. Therefore, the level of sequence similarity between the members of each cluster was analyzed, and clusters that seemed to contain two or more COGs were split.

Phylogenetic and Functional Patterns in COGs

The described analysis resulted in 710 apparent COGs. This set appears to be essentially complete as far as orthologous relationships are concerned. Indeed, when the portion of the database of proteins from complete genomes not included in the COGs was clustered by sequence similarity (*16*), only 10 groups were identified, which, upon careful inspection of the alignments, were considered likely to constitute additional COGs missed originally. These groups were incorporated, producing the final collection of 720 COGs, including 6814 proteins and distinct domains of multidomain proteins (6646 distinct gene products, or 37% of the total number of genes in the seven complete genomes) (*17*).

Most of the COGs are relatively small groups of proteins. One-third of the COGs (240 COGs with 1406 proteins) contain one representative of each of the included species (no paralogs), and 192 more COGs include paralogs from only one species, most frequently yeast (87 COGs). The mean number of proteins per COG increases with increasing number of genes in a genome, from 1.2 for *M. genitalium* to 2.9 for yeast. A notable aspect of many COGs is the differential behavior of paralogs. It is typical that one of the paralogs, for example, in yeast, shows consistently higher similarity to the orthologs in all or most of the other species (Fig. 1, B and C). For numerous yeast paralogs, particularly components of the translation apparatus, the underlying cause is obvious: the gene whose product is most similar to the bacterial orthologs is of mitochondrial origin (Fig. 1B). A more common explanation for the asymmetry of the relationships in the COGs, however, is that the highly conserved paralog has retained the original function, whereas the functions of the less conserved paralogs have changed in the course of evolution. In the already considered example (Fig. 1C), the symmetrical component of the graph (solid lines) delineates the conserved function of the σ70 subunit of the RNA polymerase (*E. coli* RpoD), which is required for the transcription of the bulk of bacterial genes, whereas the asymmetrical BeTs (broken lines) are observed for σ subunits (*E. coli* RpoH, RpoS, and FliA) involved in the transcription of specialized gene subsets (*18*). This phenomenon appears to be widespread, as we found 549 proteins in 302

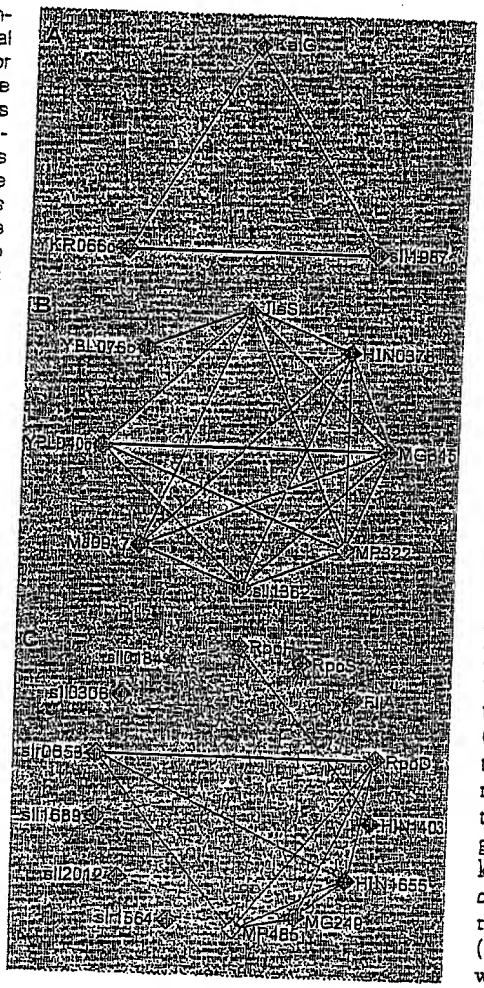


Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two yeast paralogs. Origin of the proteins: IleS, *E. coli*; HIN0378, *H. influenzae*; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; and YBL076c and YPL040c, *S. cerevisiae*. Note the adjacent triangles with a common side, for example, IleS-MG345-MJ0947 and sll1362-MG345-MJ1362. YPL040c is the yeast mitochondrial isoleucyl-tRNA synthetase; the bacterial orthologs and that from *M. jannaschii* are the BeTs for this yeast protein, but the reverse is true only of the bacterial proteins (symmetrical BeTs). Conversely, for YBL076c, which is the yeast cytoplasmic isoleucyl-tRNA synthetase, the *M. jannaschii* ortholog is a symmetrical BeT, whereas the bacterial BeTs are asymmetrical. (**C**) A complex COG with multiple paralogs. Origin of the proteins: RpoH, RpoS, RpoD, and FliA, *E. coli*; HIN1403 and HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, and slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, and MG249 are major sigma factors (σ70), whose function is universal in bacteria; note the fully symmetrical relationships between these proteins. The other proteins are specialized sigma factors whose radiation from the ancestral family apparently was accompanied by modification of the function and involved accelerated evolution; note the asymmetrical BeTs.

COGs whose corresponding paralogs showed consistently lower similarity to other members of the COG. One may think of the rapidly evolving paralogs as progenitors of new families emerging from within the conserved ones. The COGs will be an important resource in a systematic survey of the functional diversification of paralogs in conserved gene families.

There are several large clusters in the current collection with complex relationships between members. Two of these, namely the adenosine triphosphatase (ATPase) components of ABC transporters and histidine kinases, each include over 100 members. It is likely that subsequent detailed analysis of these large groups (for example, by phylogenetic tree methods) will result in their split into several distinct COGs, especially when more genomes are available. On a more general note, COGs do not supplant traditional methods of phylogenetic analysis but rather provide the appropriate starting material for these methods, in particular for a systematic analysis of phylogenetic tree topology.

Figure 2 shows the breakdown of the COGs by broadly defined function (*19*) and by species (*20*). For the majority of the COGs, the protein function is either known from direct experiments, mainly in *E. coli* or yeast, or can be confidently inferred on the basis of significant sequence similarity to functionally characterized proteins from other species. It has to be emphasized that construction of the COGs includes automatic prediction of the function for numerous genes, particularly from the poorly characterized genomes such as *M. jannaschii*. There is, however, a substantial fraction of the COGs (14%) for which only general functional prediction, typically of biochemical activity, but not the actual cellular role could be made, and for another 5%, there was no functional clue (Fig. 3). Each of the COGs includes proteins from at least three major clades whose divergence time is estimated to be over a billion years (*21*), that is, they all are ancient, conserved families with important, if not necessarily essential, cellular functions. Therefore, the proteins belonging to the "mysterious" COGs are good candidates for directed experimental studies.

The distribution of proteins from different species in the COGs shows several trends (Fig. 2), although the bias in the current collection of complete genomes (in particular, because three lineages are required to form a COG, all COGs had to have a bacterial member) must be taken into account when interpreting these comparisons. The fraction of proteins belonging to COGs is greatest in the nearly minimal genomes of mycoplasmas (70% for *M. genitalium*) and much lower in the larger genomes of *E. coli* and yeast (40% and 26%, respectively), which indeed is the tendency expected of conserved families presumably associated with cellular housekeeping functions. The genes of the pathogenic bacteria (*H. influenzae* and two mycoplasmas) are essentially subsets of the two larger bacterial gene complements, *E. coli* and *Synechocystis* sp. The latter two species almost always cooccur in the COGs. The main cause of the observed congruency is likely to be the conservation of the core of ancestral bacterial genes in nonparasitic species from different major clades. Accordingly, the fact that proteins from the pathogenic bacteria are missing in many COGs most likely testifies to gene loss, which has been extensive even in this subset of highly conserved genes. The cooccurrence of *M. jannaschii* in a COG with *E. coli* or *Synechocystis* is measurably more frequent than that with yeast (Fig. 2). Such a distribution of the archaeal genes appears to be due primarily to the blending of bacterial-like and eukaryotic-like genes in the archaeal genomes (*10*), although the mentioned bias in the genome collection is also a factor.

The phylogenetic distribution of the COG members is distinct for different functional classes (Fig. 2). It is not unexpected that translation is the only category in which ubiquitous COGs are predominant. Another obvious trend is the absence of proteins from pathogenic bacteria (*H. influenzae* and, particularly, the mycoplasmas) in many COGs

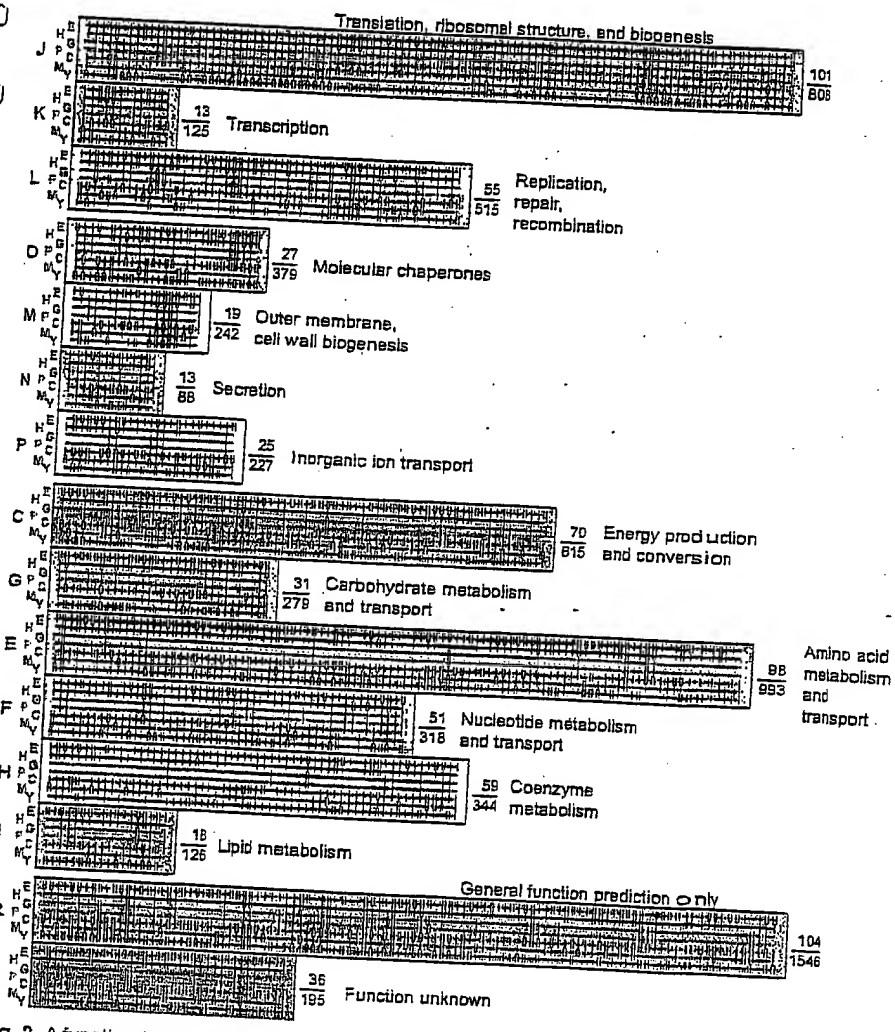


Fig. 2. A functional and phylogenetic breakdown of the COGs. E indicates *E. coli*; H, *H. influenzae*; G, *M. genitalium*; P, *M. pneumoniae*; C, *Synechocystis* sp.; M, *M. jannaschii*; and Y, *S. cerevisiae*. Each column shows a COG; a double streak indicates that two or more paralogs from the given species belong to the particular COG. The number of COGs (numerator) and the number of proteins in them (denominator) is indicated for each functional category. Capital letters in the leftmost field encode the functional categories (used in the COG IDs).

in each functional category other than translation and transcription, but especially in the metabolic functional classes. Conversely, the congruence between the two nonparasitic bacteria, *E. coli* and *Synechocystis* sp., holds for all functional classes (Fig. 2). Also apparent is the differential appearance of archaeal proteins that tend to group with yeast proteins in the translation and transcription classes (which, given the bias in the genome collection, results in ubiquitous COGs) but in all other functional classes are frequently found in COGs with bacterial proteins only.

The phylogenetic distribution of COG membership can be conveniently presented in terms of "phylogenetic patterns," which show the presence or absence of each analyzed species (Fig. 3). Of the 88 patterns that include at least three lineages (the definition of a COG), 36 were actually found. Missing were mostly patterns with only one of the two species of *Mycoplasma*, which was predictable because the gene complement of *M. genitalium* is essentially a subset of the *M. pneumoniae* complement (*22*). The remaining eight patterns that were never observed all include pathogenic bacteria without *E. coli*, which is the largest and most diverse of the available bacterial genomes. The two most abundant patterns could easily be predicted: all species ("ehgpcmy"), and all species except for the mycoplasmas ("eh__cmy"). What appears much less trivial is that these patterns together encompass only one-third of all COGs. This fact emphasizes the remarkable fluidity of genomes in evolution, revealed in spite of the fact that the analysis concentrated on ancient conserved families. Multiple solutions for the same important cellular function appear to be a rule rather than an exception, at least when phylogenetically distant species are considered (*10*, *23*). On the other hand, the eight most frequent patterns, which together account for 85% of the COGs, all include both *E. coli* and *Synechocystis*, emphasizing the congruency between these genomes.

The 114 ubiquitous COGs, most of them including components of the translation and transcription machinery, form the universal core of life. This set is more than twofold down from the bacterial "minimal set" consisting of 256 genes (*23*), but significant further erosion seems unlikely, given the broad spectrum of compared genomes.

The higher order distribution of the COGs by the three domains of life, with only 45% of the COGs including representatives of Bacteria, Archaea, and Eukarya, is another manifestation of the dynamics of gene families in evolution (Fig. 3). The picture is expected to become even more complex, and the fraction of three-domain COGs will probably drop, once archaeal-only, eukaryotic-only, and archaeal-and-eukaryotic COGs emerge with the accumulation of genome sequences.

The unusual, rare patterns are of particular interest, suggesting the possibility of unexpected findings. Each of the COGs with patterns that occur only once in our current collection (Table 1) should correspond to a unique function scattered over disconnected branches of the tree of life. Why such functions are conserved and are presumably important for survival in some but not other lineages is a challenge to be addressed experimentally. The principal evolutionary mechanisms that can be invoked to explain the emergence of these rare patterns are differential gene loss and horizontal transfer of genes. Some of the functions involved, for example, lipoate-protein ligase and glycyl–transfer ribonuclease (tRNA) synthetase, appear to be strictly essential, but in different species, they are performed by two distinct sets of orthologs unrelated to one another (*24*). Other functions, for example, thymidine phosphorylase and hexuronate dehydrogenases, may be dispensable under most conditions, and accordingly, differential gene loss is likely; it is remarkable, however, that these functions are preserved in the nearly minimal gene complements of the mycoplasmas. Two of the unique patterns, namely "__gpc_y" and "_hgp__y," might have evolved through horizontal transfer of typical eukaryotic genes into bacterial genomes. The latter pattern is of particular interest as it involves the choline kinase gene common to a number of bacterial pathogens and implicated in pathogenicity (*25*). Two of the COGs with unique patterns, "h__c_y" and "e_gp_my," include highly conserved but uncharacterized proteins whose functions could be predicted only by detailed analysis of conserved protein motifs (Table 1). These examples demonstrate the potential for protein function prediction inherent in the construction of the COGs themselves.

The sampling of genomes we compared is small and biased, and when a more complete set is available, the distribution of COGs by phylogenetic patterns is likely to change significantly; for example, many patterns that are currently rare may become common when larger genomes from the Gram-positive bacterial lineage (such as *Bacillus subtilis*) become available. Nevertheless, we believe that the language of phylogenetic patterns will become even more useful for the description of relationships between multiple genomes.

Connecting and Expanding the COGs

Ancient families of paralogs that span a broad range of taxa are well known (*26*). Accordingly, a number of COGs are related to each other and can be connected into superfamilies. In order to elucidate the superfamily structure of the COG collection, we used the recently developed PSI-BLAST (position-specific iterative BLAST) program, which combines BLAST search with profile analysis (*27*). Two COGs were considered connected if at least two of the proteins from the first COG hit members of the second COG in the PSI-BLAST search, and vice versa. Clustering by this criterion produced 58 superfamilies including 280 COGs.

Compared to COGs themselves, the superfamilies are a higher level of protein classification. Typically, they include conserved motifs that are determinants of a distinct biochemical activity, which, however, may be required for a variety of cellular functions. For example, the largest superfamily contains 53 COGs with 863 proteins, all of which contain conserved motifs typical of ATPases and GTPases but are involved in a broad range of processes from DNA replication to metabolite transport (*28*).

Superfamilies and their signature motifs

Bacteria+Eukarya +Archaea		Bacteria+Eukarya		Bacteria+Archaea		Bacteria only	
Pattern	COGs	Pattern	COGs	Pattern	COGs	Pattern	COGs
e___cmy	37	[illegible]	[illegible]	ehgpcm_	15	e_gpc__	5
eh___my	18	ehgp__y	5	e_gpcm_	4	eh_pc__	2
____cmy	13	e_gpc_y	2	_h__cm_	3		
e____my	7	e__p__y	1	eh_p_m_	2		
__gpcmy	4	e_gp__y	1	ehgp_m_	2		
_h___my	2	eh_pc_y	1	e_gp_m_	1		
eh_p_my	2	_h__c_y	1				
ehgp_my	2	___gpc_y	1				
e_gpcmy	2	_hgp__y	1				
___gp_my	1						
e_gp_my	1						
eh_pcmy	1						
Sum	323	215		122		60	
COGs(%)	45	30		17		8	

Fig. 3. Phylogenetic patterns in COGs. Letter codes as in Fig. 2 (ignore case); an underline indicates absence of the respective species. Shading indicates the eight most frequent patterns.

will be useful in classifying proteins that have evolved to an extent that they can not be assigned to any COG but still retain a conserved motif. We sought to detect such proteins with distant, subtle similarity to COGs that might be encoded in the analyzed genomes. The PSI-BLAST analysis (*27*) detected "tails" of distantly related proteins (a total of 3686) for 321 COGs, increasing the total number of proteins connected to COGs to 10,332 (58% of the entire protein set from complete genomes).

Because apparent orthologs from at least three major clades were required to form a COG, there are potential new COGs hidden among the results of the comparison of protein sequences from complete genomes (*11*). Clustering by sequence similarity the proteins not included in COGs (*14*) resulted in 443 groups with members from two clades. Predictably, the greatest number, 204, were from the cyanobacterial and Gram-negative clades, followed by 67 groups combining yeast and *M. jannaschii*. Many of these groups are likely to become COGs once additional genomes are included in the analysis.

Prediction of Protein Functions with the COG System

The COG system allows automatic functional and phylogenetic annotation of genes and gene sets (*29*). As in the procedure used for the construction of the COGs, the criterion for adding likely orthologs from other genomes to the COGs is based on the consistency between the observed relationships. A protein is compared to the database of protein sequences from complete genomes (*11*) and is included in a COG if at least two BeTs fall into it. Given that the COGs were constructed from proteins encoded in complete genomes, it is not a requirement that newly included proteins also originate from a complete genome. Indeed, while the unsequenced portion of a genome may encode proteins with the highest similarity to those included in COGs, the BeTs will not change for the products of already sequenced genes.

As a demonstration of the principle coupled with additional characterization of the COGs themselves, the sequences of proteins with known three-dimensional structures from the PDB database (*30*) were compared to the protein sequences encoded in complete genomes. The "two BeT" procedure resulted in proteins with known three-dimensional structure being included in 183 COGs, of which one was shown to be a false positive by subsequent alignment analysis. Thus, structural information could be inferred for at least 25% of the COGs. In most cases, the structurally characterized protein (from *E. coli* or yeast) actually belongs to a COG or is a closely related homolog of the proteins forming a COG.

Some of the predictions, however, provide significant functional and structural inferences. Of particular interest are (i) the possibility of modeling the nuclease domain of polyadenylate cleavage factors

Table 1. Unique phylogenetic patterns among COGs. The pattern designations are as in Fig. 3; each COG ID includes a letter indicating the functional category, to which the constituent proteins belong (Fig. 2).

Pattern and COG ID	Proteins	Activity or function	Comment
e_gp_m_ COG0213F	DeoA-MG051-MP090-MJ0667	Thymidine phosphorylase; salvage of deoxypyrimidines	Nonessential gene in *E. coli*; apparent orthologs found in other Gram-positive bacteria and in humans (*35*).
e__p__y COG0246G	MtlD, UxaB, UxuB, YdfI, YeiQ-MP190-YEL070w, YNR073c	Mannitol-1-phosphate and other hexuronate dehydrogenases; hexuronate catabolism	Nonessential genes in *E. coli*; accessory reactions of carbohydrate metabolism (*36*).
e_gp__y COG0095H	LplA-MG270-MP450-(sll0809)-YJL046w	Lipoate-protein ligase A; ligation of lipoate to apoproteins of pyruvate dehydrogenase and other lipoate-dependent enzymes	There are two unrelated classes of lipoate-protein ligases; *E. coli* and yeast encode both forms; *H. influenzae* and *Synechocystis* sp. encode the B form (included in a separate COG); sll0809 is a distant homolog of the A form (*37*), which was not automatically included in the COG but was detected with PSI-BLAST.
eh_pc_y COG0604R	AdhC + 18 *E. coli* proteins–MP278-sll0990, slr1192-YBR046c + 19 yeast proteins	Alcohol dehydrogenase class III and related Fe-S dehydrogenases; various catabolic pathways	Highly conserved protein family distinct from other Fe-S oxidoreductases.
_h__c_y COG0678R	HIN1693_1-sll1621-YLR109w	Glutaredoxin-like membrane protein (prediction)	The *H. influenzae* protein contains an additional thioredoxin-like domain.
__gpc_y COG0631R	MG108-MP586-sll1771-sll1033-sll0602-YDL006w + 6 yeast proteins	Protein serine and threonine phosphatase	Serine and threonine protein phosphatases are abundant in eukaryotes but not in bacteria (*38*).
__gp_my COG0423J	MG251-MP483-MJ0228-YPR081c, YBR121c	Glycyl-tRNA synthetase (eukaryotic and Gram-positive type)	Gram-negative bacteria and *Synechocystis* encode a distinct glycyl-tRNA that appears to be unrelated to the eukaryotic and Gram-positive type; the closest relative of this COG in *E. coli* and *H. influenzae* is prolyl-tRNA synthetase (*24*).
e_gp_my COG0622R	b2300-MG207, MP029-MJ0623, MJ0936-YHR012w	Phosphoesterase (prediction)	Highly conserved protein family that shares only modified catalytic motifs (detected by PSI-BLAST; $P \sim 0.004$) with other phosphoesterases, including protein phosphatases.
eh_pcmy COG0078E	ArgI, ArgF, YgeW-HIN0012-MP531-sll0902-MJ0881-YJL088w	Ornithine carbamoyltransferase; arginine biosynthesis	Amino acid metabolism appears to be completely missing in *M. genitalium*, but residual reactions may occur in *M. pneumoniae*.
_hgp_y COG0510M*	HIN0938-MG356, MP310-YDR147w, YLR133w	Choline kinase (prediction) involved in lipopolysaccharide biosynthesis	Enzyme common to several bacterial pathogens and eukaryotes; contributes to pathogenicity (*25*).

*This COG was added to the collection by cluster analysis.

(31) with the beta-lactamase structure, (ii) the presence of an acylphosphatase domain in hydrogenase expression factors, which form a highly conserved COG, and in a number of uncharacterized proteins, and (iii) the connection between a unique carbonic anhydrase and an acetyltransferase family (Table 2).

Probably the most important application of the COGs is functional characterization of newly sequenced genomes. In the preliminary analysis of the recently published genome of the major human bacterial pathogen *Helicobacter pylori* (32), 813 proteins (51% of the gene products) from this bacterium were included in 453 preexisting COGs and 143 new COGs (33). In spite of the fact that many *H. pylori* proteins are highly similar to homologs from *E. coli* and other bacteria and have been explored in detail (32), this analysis produced over 100 additional functional predictions (33).

Conclusions and Perspective

The COGs bring together the fields of comparative genomics and protein classification. Among the numerous possible approaches to protein classification, the COGs appear to be unique as a prototype of a natural system, which has as its basic unit a group of descendants of a single ancestral gene. Typically, such a group is associated with a conserved, specific function, so that the inclusion of a protein in a COG automatically entails functional prediction.

Each COG contains conserved genes from at least three phylogenetically distant clades and, accordingly, corresponds to an ancient conserved region (ACR). Previous analyses have indicated that the total number of distinct ACRs is likely to be less than 1000 (34). Thus, even with the limited number of complete genomes currently available for analysis, the COGs have already captured a substantial fraction of all existing highly conserved protein domains. With more genomes included in the system, the discovery of additional COGs should gradually level off, with the great majority of the ACRs encoded in the added genomes fitting into already known COGs.

With the forthcoming flood of genome sequences, a coherent framework for understanding these genomes from both the functional and evolutionary viewpoints is a must. We regard the current collection of

Table 2. Structural and functional predictions for uncharacterized proteins in COGs.

Phylogenetic pattern and COG ID*	Proteins in COG†	Activity and function	Homolog in PDB‡ •BeTs detected (no.) •Lowest *P* with a COG member	Comment
e_gpcmy COG0595R	PhnP, ElaC-2g-2p-5c-8m- YLR277c, YMR137c, YKR079c	Predicted Zn-dependent hydrolases	Beta-lactamase (1BMC) •2 •0.039	Activity is not known for any protein in this ubiquitous COG. Biochemical and genetic data indicate that YLR277c is involved in messenger RNA 3′-end processing (*31*), whereas YMR137c is DNA cross-link repair protein SNM1 (*39*). A motif including the Zn-coordinating histidines of beta-lactamase is conserved.
eh_.cmy COG0607R	SseA, PspE, GlpE, YibN, YbbB, YnjE, YgaP-2h-5c-MJ0052-4y	Predicted sulfur- transferases	Rhodanese (1RHD, 2ORA, 1ORB) •2 •10^{-41}	The sulfurtransferase activity of SseA has been demonstrated (*40*), but the rest of the proteins in this COG have no known activity. PspE (phage shock protein), GlpE (uncharacterized protein involved in glycerol metabolism), and other small proteins correspond to one of the two rhodanese domains.
ehgpc_y COG0596R	PldB, MhpC, YcdJ, YnbC-HIN0065- MG020-MP132–6c- YNR064c, YKL094w	Predicted hydrolases and acyltransferases	Lipases (2LIP, 1TAH|B, 1CVL) •3 •8×10^{-5}	PldB is known to possess triglyceride lipase activity (*41*). All other proteins in the COG have not been characterized but now can be predicted to possess the α- or β-hydrolase fold.
e__cm_ COG0068C	HypF-sll0322-MJ0713	Hydrogenase maturation factor	Acylphosphatase (1APS) •2 •2×10^{-5}	HypF is required for hydrogenase biosynthesis (*42*), but no biochemical activity is known. The ~100 amino acid, NH_2-terminal domain aligns with acylphosphatase, with the catalytic residues conserved, suggesting that HypF orthologs indeed possess acylphosphatase activity. A PSI-BLAST search with this domain as the query detected five additional likely acylphosphatases, namely *E. coli* YccX and *M. jannaschii* MJ0809, MJ0553, MJ1331, and MJ1405 (*43*).
e__cm_ COG0663R	CaiE, YrdA, YdbZ-sll1636, sll1031-MJ0304	Predicted carbonic anhydrases	Carbonic anhydrase from Methanosarcina thermophila (1THJ) •3 •10^{-29}	The biochemical activity of the proteins in this COG is not known. They show not only conservation of histidine residue comprising the active center of this unusual carbonic anhydrase (*44*) but also significant similarity to acetyltransferases of the isoleucine patch superfamily (*45*), suggesting an unexpected connection between the two types of enzymes.

*The designations are as in Table 1 and Fig. 3. †2g indicates two proteins from *M. genitalium*, 2p indicates two proteins from *M. pneumoniae*, and so forth. ‡The PDB accession is indicated in parentheses.

COGs as a crude first version of such a framework. Inclusion of additional, phyloÛ genetically diverse genomes and further deÛ velopment of the procedures used to derive and analyze COGs will hopefully result in refinement of this system, making it a solid platform for genome annotation and evoluÛ tionary genomics.

REFERENCES AND NOTES

1. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
2. C. M. Fraser *et al.*, *ibid.* 270, 397 (1995); R. Himmelreich *et al.*, *Nucleic Acids Res.* 24, 4420 (1996); T. Kaneko *et al.*, *DNA Res.* 3, 109 (1996); F. R. Blattner *et al.*, *Science* 277, 1453 (1997).
3. C. J. Bult *et al.*, *Science* 273, 1058 (1996).
4. A. Goffeau *et al.*, *ibid.* 274, 546 (1996); H. W. Mewes *et al.*, *Nature* 387, 7 (1997).
5. C. R. Woese, *Curr. Biol.* 6, 1060 (1996); G. J. Olsen and C. R. Woese, *Cell* 89, 991 (1997); E. V. Koonin, *Genome Res.* 7, 418 (1997).
6. E. V. Koonin, A. R. Mushegian, K. E. Rudd, *Curr. Biol.* 6, 404 (1996); E. V. Koonin and A. R. Mushegian, *Curr. Opin. Genet. Dev.* 6, 757 (1996).
7. W. M. Fitch, *Syst. Zool.* 19, 99 (1970). This definition may not embrace all of the complexity of relationships between genes in different genomes. For example, if genes A and B are paralogs encoded in genome 1, and A′ and B′ are their respective orthologs in genome 2, what is the appropriate description of the relationship between A and B′? They formally are not paralogs, even though a generalized definition might include such cases. Furthermore, one-to-many and many-to-many orthologous relationships evidently exist.
8. W. M. Fitch, *Philos. Trans. R. Soc. London Ser. B* 349, 93 (1995).
9. R. L. Tatusov *et al.*, *Curr. Biol.* 6, 279 (1996).
10. E. V. Koonin, A. R. Mushegian, M. Y. Galperin, D. R. Walker, *Mol. Microbiol.* 25, 619 (1997).
11. The protein sequences were from the original references (*1*–*4*), with modifications (for example, tentative correction of frame-shift errors) and additions (previously unreported predicted genes) made for *E. coli* (E. V. Koonin and R. L. Tatusov, unpublished observations; K. E. Rudd, personal communication), *H. influenzae* (*9*), *M. genitalium* and *M. jannaschii* (*10*), and *S. cerevisiae* (T. J. Wolfsberg and D. Landsman, personal communication). The list of systematic names for all *E. coli* genes was provided by K. Rudd, and the names for all yeast genes were provided by T. Wolfsberg and D. Landsman; the *H. influenzae* genes were renamed as previously described (*9*); the gene names for the other species were from the original publications. The resulting protein database from complete genomes used in all comparisons contained 4283 sequences from *E. coli*, 1703 sequences from *H. influenzae*, 468 sequences from *M. genitalium*, 677 sequences from *M. pneumoniae*, 3168 sequences from *Synechocystis* sp., 1736 sequences from *M. jannaschii*, and 5932 sequences from *S. cerevisiae*, totaling 17,967 sequences. This sequence set is available on the World Wide Web at http://www.ncbi.nlm.nih.gov/COG. All pairwise comparisons between these sequences were performed using the BLASTPGP program, which is based on an enhanced version of the BLAST algorithm and includes analysis of local alignments with gaps (*26*). Predicted coiled coil regions in protein sequences were masked before the comparison using the batch version of the COILS2 program [A. Lupas, *Methods Enzymol.* 266, 513 (1996); D. R. Walker and E. V. Koonin, *ISMB* 5, 333 (1997)], and additionally, regions of low complexity were masked using the SEG program with default parameters [J. C. Wootton and S. Federhen, *Methods Enzymol.* 266, 554 (1996)]. Before the detection of triangles of BeTs, paralogs were identified as those proteins from the same lineage that showed greater similarity to each other than to any protein from another lineage. For the purpose of triangle formation, paralogs were treated as a group. The algorithm further included verification that the BeTs included in a triangle formed a consistent multiple alignment; triangles that did not contain a conserved motif were disregarded.
12. Although the exact solution depends on the amino acid composition and size of the particular proteins, under zero approximation, if B (from genome *b*) is the BeT for A (from genome *a*), and C (from genome *c*) is the BeT for B, the probability that C is the BeT for A by chance is close to 1/*N*, where *N* is the number of genes in genome *c*, or ~0.001.
13. C. R. Woese, *Microbiol. Rev.* 51, 221 (1987); ———, R. Overbeek, G. J. Olsen, *J. Bacteriol.* 176, 1 (1994); N. R. Pace, *Science* 276, 734 (1997). A BeT to a given clade was registered if detected in any of the constituent species, for example, in *E. coli* or *H. influenzae* for the Gram-negative bacteria.
14. H. Watanabe and J. Otsuka, *Comput. Appl. Biosci.* 11, 159 (1995); E. V. Koonin, R. L. Tatusov, K. E. Rudd, *Methods Enzymol.* 266, 295 (1996).
15. A schematic visual representation of the search results was used for this analysis [T. L. Madden, R. L. Tatusov, J. Zhang, *Methods Enzymol.* 266, 131 (1996)].
16. A single-linkage clustering procedure was used with random match probability, $P < 0.001$, as the cutoff (*14*).
17. A searchable database of COGs is available at http://www.ncbi.nlm.nih.gov/COG. Each COG was assigned a unique identification number, which includes a letter for the functional category (*19*) and a number (see examples in Fig. 1 and Tables 1 and 2).
18. M. Lonetto, M. Gribskov, C. A. Gross, *J. Bacteriol.* 174, 3843 (1992).
19. The broad functional categories of proteins were as defined previously (*9*), except that transcription was separated from replication, recombination, and repair. This classification is a modification of the system originally developed for *E. coli* proteins [M. Riley, *Microbiol. Rev.* 57, 862 (1993)].
20. A partially similar representation of some of the protein families from complete genomes has been recently published [R. A. Clayton, O. White, K. A. Ketchum, J. C. Venter, *Nature* 387, 459 (1997)].
21. R. F. Doolittle, D.-F. Feng, S. Tsang, G. Chao, E. Little, *Science* 271, 470 (1996).
22. R. Himmelreich *et al.*, *Nucleic Acids Res.* 25, 701 (1997).
23. A. R. Mushegian and E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* 93, 10268 (1996).
24. E. V. Koonin, A. R. Mushegian, P. Bork, *Trends Genet.* 12, 334 (1996).
25. J. N. Weiser, M. Shchepetov, S. T. Chong, *Infect. Immun.* 65, 943 (1997).
26. J. P. Gogarten *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 86, 6661 (1989); N. Iwabe *et al.*, *ibid.*, p. 9355; J. P. Gogarten, E. Hilario, L. Olendzewski, in *Evolution of Microbial Life*, D. McL. Roberts, P. Sharp, G. Alderson, M. Collins, Eds. (Cambridge Univ. Press, Cambridge, 1996), pp. 267–292.
27. S. F. Altschul *et al.*, *Nucleic Acids Res.* 25, 3389 (1997). The probability of a random match, $P < 0.001$, was used in all PSI-BLAST searches.
28. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, *EMBO J.* 1, 945 (1982); A. E. Gorbalenya and E. V. Koonin, *Nucleic Acids Res.* 17, 8413 (1989); M. Saraste, P. R. Sibbald, A. Wittinghofer, *Trends Biochem. Sci.* 15, 430 (1990).
29. Protein sequences can be submitted for searching against COGs at http://www.ncbi.nlm.nih.gov/COG/cognitor.html
30. F. C. Bernstein *et al.*, *J. Mol. Biol.* 112, 535 (1977).
31. G. Chantreau, S. M. Noble, C. Guthrie, *Science* 274, 1511 (1996); A. Jenny, L. Minvielle-Sebastia, P. J. Preker, W. Keller, *ibid.* 274, 1514 (1996); G. Stumpf and H. Domdey, *ibid.*, p. 1517.
32. J.-F. Tomb *et al.*, *Nature* 388, 539 (1997).
33. E. V. Koonin, R. L. Tatusov, M. Y. Galperin, M. N. Rozanov, unpublished observations.
34. P. Green *et al.*, *Science* 259, 1711 (1993).
35. J. Neuhard and R. A. Kelln, in Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, ed. 2, 1996), pp. 580–599.
36. E. C. C. Lin, *ibid.*, pp. 307–342.
37. T. W. Morris, K. E. Reed, J. E. Cronan Jr., *J. Bacteriol.* 177, 1 (1995).
38. P. Bork, N. P. Brown, H. Hegyi, J. Schultz, *Protein Sci.* 5, 1421 (1996).
39. D. Richter, E. Niegemann, M. Brendel, *Mol. Gen. Genet.* 231, 194 (1992); R. Wolter, W. Siede, M. Brendel, *ibid.* 250, 162 (1996).
40. H. Hama, T. Kayahara, W. Ogawa, M. Tsuda, T. Tsuchiya, *J. Biochem.* 115, 1135 (1994).
41. T. Kobayashi *et al.*, *ibid.* 98, 101 (1985).
42. A. Colbeau *et al.*, *Mol. Microbiol.* 8,15 (1993).
43. M. N. Rozanov and E. V. Koonin, unpublished observations.
44. B. E. Alber and J. G. Ferry, *Proc. Natl. Acad. Sci. U.S.A.* 91, 6909 (1994); C. Kisker *et al.*, *EMBO J.* 15, 2323 (1996).
45. E. V. Koonin, *Protein Sci.* 4, 1608 (1995); M. N. Rozanov and E. V. Koonin, unpublished observations.

46. We thank A. Schaffer for modifying the PSI-BLAST program; R. Walker, H. Watanabe, and M. Rozanov for valuable help with data analysis; K. Rudd, T. Wolfsberg, and D. Landsman for unpublished data; and P. Bork, M. Galperin, M. Gelfand, A. Mushegian, P. Pevzner, M. Roytberg, M. Rozanov, and R. Walker for helpful discussions.

Microbial pathogen genomes – new strategies for identifying therapeutics and vaccine targets

Douglas R. Smith

Advances in high-throughput DNA-sequencing techniques have given us the unprecedented ability to rapidly determine the nucleotide sequences of entire bacterial genomes. The application of these methods to the genomes of microbial pathogens, combined with efficient analytical tools and genome-scale approaches for studying gene expression, is revolutionizing our approach to the selection of targets for drug screening and vaccine development. This is bringing new life to this important, but long-neglected, field of research.

The decision, several years ago, by the US Department of Energy, the National Institutes of Health (NIH) and several international funding agencies to embark upon programs to map and sequence the human genome has led to a number of important technological advances that are beginning to have an impact in other areas of biology. Among these advances are the development of automated methods for the generation of large amounts of raw DNA-sequencing information, computer software for rapidly processing and analyzing primary sequence data, and techniques for the rapid assembly of shotgun sequencing reads, even from entire bacterial genomes. Efficient algorithms for similarity searching allow the rapid identification of protein-encoding sequences that are homologous to other genes, the sequences of which are held in public and private databases; as from April 1996, approximately 500 megabases (Mb) of nucleotide sequence were contained in GenBank, and approximately 200 000 sequences were held in the SWISS-PROT/Genpept/PIR database of non-redundant proteins. Combined with the wealth of biochemical information that is archived in public databases, it has become possible to describe rapidly the full repertoire of genes in a microbial genome, and to predict many of the metabolic pathways that an organism may utilize.

Progress in this field has been stimulated by the interests of the biotechnology and pharmaceutical industries in using genome-sequencing data as a basis for drug discovery. In turn, this has led to the development of proprietary databases containing genomic information, which provide the basis for *in silico* experiments to identify novel targets for drugs, and for laboratory experiments to identify genes that perform critical functions. This article summarizes some recent developments in this important area, focusing on bacterial sequences, and provides examples to illustrate how genome-sequencing information from microbial pathogens can be used to select targets for vaccine and drug development. The overall process used to proceed from sequence generation to target validation is illustrated in Fig. 1.

D. R. Smith (smith@tric.com) is at Genome Therapeutics Corporation, 100 Beaver Street, Waltham, MA 02154, USA.

Large-scale sequencing of bacterial genomes

Many laboratories use automated sample-preparation techniques and fluorescence-based gel readers [such as that produced by Applied Biosystems Inc., (ABI); Foster City, CA, USA] for the large-scale sequencing of bacterial genomes. These instruments have the advantage that they are efficient, and relatively easy to set up and operate. A few laboratories use computer-assisted multiplex sequencing to achieve the same end[1]. In multiplex sequencing, samples consisting of pools of up to 20 plasmids are processed through sample preparation and gel electrophoresis, and the resulting sequences are determined from electroblots of the gels by hybridization with radioactive or fluorescently labeled probes. This technique can be used to generate 40 films (or digitized images) from each sequencing gel. Although multiplex sequencing is efficient at producing large amounts of 'shotgun' data, it is more difficult to set up and operate in the laboratory than is fluorescence-based gel sequencing, and it is not suited to directed-finishing strategies. ABI machines are used in the author's laboratory to generate primer-directed reads for finishing and gap closure.

During the past year, a group at The Institute for Genomic Research (TIGR; Gaithersburg, MD, USA) reported the complete sequences of *Haemophilus*

 0167 – 7799/96/$15.00. PII: S0167-7799(96)10038-X

EXHIBIT 3

influenzae (1.8 Mb), a major cause of respiratory infections and meningitis, especially in children[2], and of *Mycoplasma genitalium* (0.6 Mb), which causes urethritis[3]. Approximately 1.6 Mb of contiguous sequence from the 4.7 Mb *Escherichia coli* genome has been published[4], and the sequencing of a further 2 Mb was reported at the 1995 Genome Sequencing and Analysis VII (GSA-VII) meeting[5]. The genome of *Helicobacter pylori* (1.7 Mb), the major cause of stomach ulcers, has been sequenced by Genome Therapeutics Corporation (GTC; Waltham, MA, USA) under a privately funded microbial-pathogen sequencing program. More than half (1.5 Mb) of the 2.8 Mb genome of *Mycobacterium leprae* (the etiologic agent of leprosy) has also been sequenced by GTC, and is available through GenBank, the GTC web site <http://www.tric.com>, and through MycDB <http://www.biochem.kth.se/MycDB.html>, which contains mycobacterial genome mapping and sequence information[6].

Other microbial pathogens that are currently being sequenced include *Neisseria gonorrhoeae* (University of Oklahoma, Norman, OK, USA), *Streptococcus pyogenes* (University of Oklahoma), *Treponema pallidum* (University of Texas, Houston, TX, USA, and TIGR), *Mycobacterium tuberculosis* (GTC and the Sanger Centre, Hinxton, Cambridge, UK), and *Staphylococcus aureus* [GTC, and Human Genome Sciences (HGS; Rockville, MD, USA)].

In addition to these pathogens, the genomes of several archaebacteria and other non-pathogens are being sequenced. These include *Methanococcus janaschii* (TIGR), *Pyrococcus furiosis* (University of Utah, Salt Lake City, UT, USA), *Sulfolobus solfataricus* (Dalhousie University, Halifax, Nova Scotia, Canada), and *Pyrobaculum aerophilum* (California Institute of Technology, Pasadena, CA, USA, and University of California, Los Angeles, CA, USA). The 1.7 Mb genome of the archaeon *Methanobacterium thermoautotrophicum* is near completion at GTC (Ref. 7). Approximately 2 Mb of the 4.1 Mb *Bacillus subtilis* genome has now been sequenced by a consortium of European and Japanese laboratories, and the project may be completed by the end of 1996 (Ref. 8). Approximately 1 Mb of genomic sequence from the 2.7 Mb genome of the cyanobacterium *Synechocystis* sp. 6803 was recently published[9].

Within the next couple of years, therefore, we can expect an explosion of bacterial-genome sequence information from species representing a variety of phylogenetic lineages, including many pathogens.

Pharmaceutical companies have shown considerable interest in using pathogen genomics to facilitate the development of vaccines and small-molecule therapeutics. For example, researchers at GlaxoWellcome have sequenced a substantial fraction of the *H. pylori* genome to assist in the process of drug discovery. Over the past year, GTC has formed two research alliances with pharmaceutical companies to take advantage of sequences from microbial pathogens: one with Astra AB, focusing on the development of new antibiotics and vaccines to treat *H. pylori* infection, and one with Schering-Plough (Union, NJ, USA), to develop broad-spectrum antibiotics and vaccines. Although the genomic route to drug discovery for bacterial pathogens is new and remains unproved, the basic paradigm (outlined below) of gene identification, followed by functional analysis and drug screening, is well established. Thus, it is likely that more companies will become involved, and that in the the future, additional research alliances between genomics companies and the pharmaceutical industry will materialize in this area.

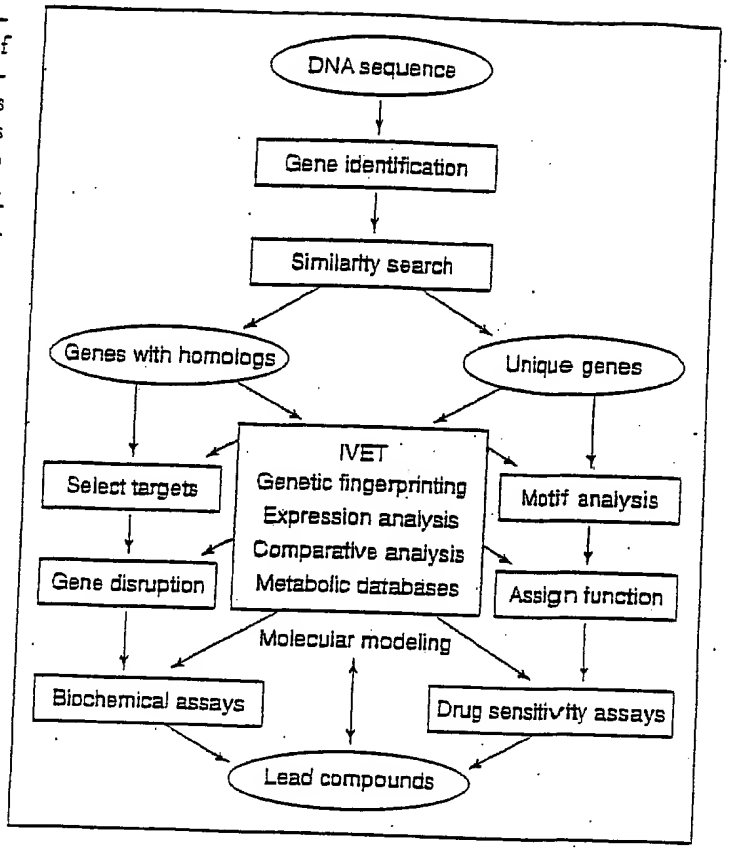
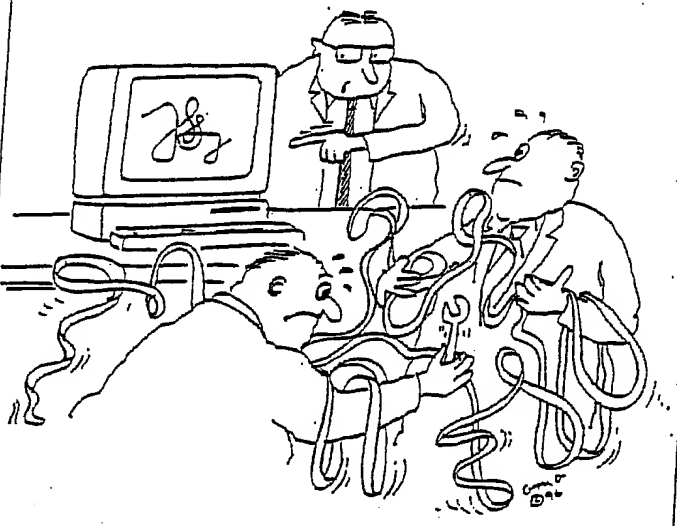


Figure 1

Flow diagram illustrating the process by which a microbial genome sequence is analysed and the information is used to direct experiments and aid in target selection for therapeutics development. The individual steps are referred to throughout the text. In the case of vaccine candidates, gene products from selected targets are expressed and tested in animal models.

From sequence to genes

The first task when confronted with an entire bacterial-genome sequence, is to identify all the genes. This can be accomplished using a variety of techniques, but the most successful approaches use a combination of reading-frame and codon-usage analysis, together with similarity searching, to identify putative genes with homology to previously described sequences. Commonly used tools include GeneMark[10], GenomeBrowser[11], BLAST (Ref. 12), and highly parallelized implementations of the Smith–Waterman



'Well, according to the algorithms, it folds like this!'

alignment, such as BLAZE, or MPsrch (Ref. 13). In general, organism-specific codon usage is highly predictive for bacterial genes, but its effective use depends on the existence of sufficient information to generate accurate codon-usage matrices. In some cases, subsets of genes within an organism will exhibit codon-usage patterns that deviate significantly from the norm[14]. Such genes are thought to represent evolutionarily recent acquisitions by phage transduction, conjugation, or some other form of horizontal transfer from other organisms. If enough of these genes are present, codon-usage tables of genomic subsets can be constructed to identify them. Translational start sites can be identified by the occurrence of start codons that coincide with abrupt changes in codon usage, the initiation of homology to previously characterized genes, or the presence of Shine–Dalgarno sequences[15]. Automated analysis tools (such as GenomeBrowser[11]) that provide a graphical display of open reading frames (ORFs), codon usage, database homologies and other features, make the task of identifying bacterial genes and their relationships with each other in the genome relatively straightforward. With the increasing pace of bacterial-genome sequencing, there is an emerging need for second-generation tools that will automate most of the laborious annotation process.

From genes to function

The second phase in the analysis of bacterial genomes is to identify the function of as many genes as possible. Currently, sequence homology is the most powerful tool. A high degree of homology between the putative translation product of a newly identified gene and an enzyme whose function has been thoroughly studied in other organisms, provides strong support for the function of that protein, especially if it is the only homolog in the genome under scrutiny. Other useful tools include programs that identify sequence motifs from databases such as PROSITE (Ref. 16), BLOCKS (Ref. 17), BEAUTY (Ref. 18) and ProDom (Ref. 19). If one is attempting to identify vaccine candidates, then examining highly expressed cell-surface proteins is relevant, so it is then useful to know whether a protein contains a secretion signal, even if nothing else is known about it. Although the tools described here are very good at identifying homologies, 25–40% of the genes in a bacterial genome typically fail to show significant similarity with known proteins.

Once the set of similarity-searching tools has been exhausted, one must return to molecular biology to further elucidate the function and expression pattern of predicted genes. Commonly used approaches to identifying essential genes in an organism include: the use of gene knockouts, disruptions using transposon-mediated mutagenesis, or homologous recombination with disrupted gene-constructs that contain an antibiotic-resistance cassette. Gene disruptions can be generated in a variety of ways, including sophisticated 'hit-and-run' approaches that interrupt a gene without introducing polar effects into downstream ORFs (Ref. 20). However, a gene-by-gene approach to the study of a whole genome is certainly time consuming and labor intensive.

The availability of large amounts of genome-sequence information has stimulated the development of new approaches to functional analysis on a genomic scale. This has been particularly true for researchers investigating yeast, where a concerted effort is being made to ascertain the function of every ORF in the genome. Such strategies include the conceptually simple, but technologically advanced, technique of making microarrays of polymerase chain reaction (PCR)-amplified gene sequences on glass slides to allow the fluorescence-based detection of quantitative hybridization signals from labeled cDNA probes on large numbers of genes simultaneously — perhaps even all the genes of an organism[21]. An ingenious PCR-based approach to efficient sequence-signature-based expression analysis has recently been demonstrated[22]. For example, a technique termed 'genetic fingerprinting' promises to replace individual gene knockouts by a global transposon-mutagenesis approach[23]. Insertions are induced *en masse* in a strain of interest, the strain is grown under a variety of conditions, and PCR products are analysed to identify genes in which transposon hops are under-represented because the genes are required for growth[23]. A conceptually similar dropout technique, which uses tagged transposons to identify the *Salmonella typhimurium* genes required for virulence in a mouse model, has been described[24].

Techniques that probe subsets of genes for a specific functionality, such as secretion or induction during growth in the host, have also been described. These techniques provide clones from which signature

sequences can be derived, so that corresponding genes can be identified by comparing them with the genomic sequence. The IVET (*in vivo* expression technology) technique, which detects gene fusions that result in the *in vivo* selectable expression of a defective *purA* gene or antibiotic-resistance marker, has been used to identify *Salmonella* genes, the expression of which is induced when the pathogen is grown in mice[25]. Finally, protein microsequencing[26] and mass-spectrometry-based peptide analysis[27] have been used to identify protein components (e.g. outer-membrane proteins) in partially purified mixtures, or to identify specific proteins separated by two-dimensional gel electrophoresis. Sequences generated in this manner can be used to correlate specific proteins with the gene sequences from which they are expressed.

Target selection and validation

The techniques described in the previous section can be used to identify genes in specific functional categories that may represent good targets for drug or vaccine development. In general, when developing new antibiotics, one is interested in genes that are essential under all growth conditions (and preferably even in quiescent cells), and for which inhibitors with useful chemical properties, such as permeability and low toxicity, can be identified. One advantage of having the entire sequence of a genome is that targets can be prioritized in terms of their activities and the properties of compounds that are known to interact with them. Even with the results of knockout or *in vivo* expression experiments, additional biological information can aid in narrowing down the field of choices. For example, genes can be selected on the basis of their probable roles in intracellular metabolism. Databases, such as EcoCyc (Ref. 28) or PUMA (Ref. 29), that describe known metabolic pathways can be helpful in this regard. Detailed structural information about homologs of identified genes (determined using the Protein DataBank[30]) can be used to assist in the molecular modeling of inhibitors (some resources for molecular modeling can be found at Ref. 31).

As more genomes are sequenced, it will become possible to identify genes that are unique to a particular organism or group of organisms, or genes that are conserved in certain groups. Thus, for example, it will be possible to use electronic comparison to identify genes that are present in *H. pylori* but not in other gut-dwelling bacteria such as *E. coli*, providing a basis for the development of antibiotics specific to *H. pylori*. Although combinatorial chemistries promise to speed up our ability to synthesize and screen large numbers of unique chemical entities, the sequence-based approach described here provides an avenue for the rational identification and selection of key targets for therapeutics development. Ultimate validation of the targets will, of course, require additional experiments such as protein expression, biochemical-assay development and animal studies to identify those with the most useful properties or inhibitors.

Acknowledgements

The sequencing of *Mycobacterium leprae* and *M. tuberculosis*, and technology development for multiplex sequencing is supported by a NIH Genome Science and Technology Center grant 1P01-HG1106-01 from the National Center for Human Genome Research. The sequencing of *Methanobacterium thermoautotrophicum* is supported under the Microbial Genome Program by Grant No. DE-FC02-95ER61967 from the Office of Health and Environmental Research of the US Department of Energy. The sequencing of *Helicobacter pylori* and *Staphylococcus aureus* is supported by Genome Therapeutics Corporation. Thanks to Brad Guild for comments on the manuscript.

References

1 Church, G. M. and Kieffer-Higgins, S. (1988) *Science* 240, 185–188
2 Fleischmann, R. D. *et al.* (1995) *Science* 269, 496–512
3 Fraser, C. D. *et al.* (1995) *Science* 270, 397–403
4 Burland, V., Plunkett, G., III, Sofia, H. J., Daniels, D. L. and Blattner, F. R. (1995) *Nucleic Acids Res.* 23, 2105–2119
5 Burland, V., Plunkett, G., III and Blattner, F. R. (1995) in *Genome Science and Technology* 1, P-16, Mary Ann Liebert
6 Bergh, S. and Cole, S. T. (1994) *Mol. Microbiol.* 12, 517–534
7 Smith, D. R. *et al.* (1995) in *Genome Science and Technology* 1, P-48, Mary Ann Liebert
8 Devine, K. (1995) *Trends Biotechnol.* 13, 210–216
9 Kaneko, T. *et al.* (1995) *DNA Res.* 2, 153–166
10 Borodovsky, M. and McInnich, J. (1993) *Comput. Chem.* 17, 123–133
11 Robison, K. R. and Church, G. M. (1995) <http://www.belmont.com/gb.html>
12 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410
13 MPsrch <http://www.ebi.ac.uk/searches/blitz.html>
14 Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) *J. Mol. Biol.* 222, 851–856
15 Shine, J. and Dalgarno, L. (1975) *Eur. J. Biochem.* 57, 221–230
16 Bairoch, A. (1991) *Nucleic Acids Res.* 19, 2241–2245
17 Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acids Res.* 19, 6565–6572
18 Worley, K. C., Wiese, B. A. and Smith, R. F. (1995) *Genome Res.* 5, 173–184
19 Sonnhammer, E. L. and Kahn, D. (1994) *Protein Sci.* 3, 482–492
20 Link, A. J. and Church, G. M. <http://twod.med.harvard.edu/labgc/pKO3.html>
21 Schena, M., Shalon, D., Davis, R. D. and Brown, P. O. (1995) *Science* 270, 467–470
22 Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) *Science* 270, 484–487
23 Smith, V., Botstein, D. and Brown, P. O. (1995) *Proc. Natl Acad. Sci. USA* 92, 6479–6483
24 Hensel, M. *et al.* (1995) *Science* 269, 400–403
25 Mahan, M. J. *et al.* (1995) *Proc. Natl Acad. Sci. USA* 92, 669–673
26 Tempst, P., Link, A. J., Riviere, L. R., Fleming, M. and Elicone, C. (1990) *Electrophoresis* 11, 537–553
27 James, P., Quadroni, M., Carafoli, E. and Gonnet, G. (1993) *Biochem. Biophys. Res. Commun.* 195, 58–64
28 Karp, P. D. (1992) *CABIOS* 8, 347–357
29 Gaasterland, T., Maltsev, N., Overbeek, R., Selkov, E. <http://www.mcs.anl.gov/home/compbio/PUMA>
30 Protein DataBank <http://www.pdb.bnl.gov>
31 <http://www.pharmacy.wisc.edu>

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 273, No. 15, Issue of April 10, pp. 9058–9069, 1998
Printed in U.S.A.

Functional and Structural Heterogeneity of the DNA Binding Site of the *Escherichia coli* Primary Replicative Helicase DnaB Protein*

(Received for publication, October 16, 1997, and in revised form, December 17, 1997)

Maria J. Jezewska, Surendran Rajendran, and Wlodzimierz Bujalowski‡

From the Department of Human Biological Chemistry and Genetics and the Sealy Center for Structural Biology, University of Texas Medical Branch at Galveston, Galveston, Texas 77555-1053

The structure-function relationship within the DNA binding site of the *Escherichia coli* replicative helicase DnaB protein was studied using nuclease digestion, quantitative fluorescence titration, centrifugation, and fluorescence energy transfer techniques. Nuclease digestion of the enzyme-single-stranded DNA (ssDNA) complexes reveals large structural heterogeneity within the binding site. The total site is built of two subsites differing in structure and affinity, although both occlude ~10 nucleotides. ssDNA affinity for the strong subsite is ~3 orders of magnitude higher than that for the weak subsite.

Fluorescence energy transfer experiments provide direct proof that the DnaB hexamer binds ssDNA in a single orientation, with respect to the polarity of the sugar-phosphate backbone. This is the first evidence of directional binding to ssDNA of a hexameric helicase in solution. The strong binding subsite is close to the small 12-kDa domains of the DnaB hexamer and occludes the 5′-end of the ssDNA. The strict orientation of the helicase on ssDNA indicates that, when the enzyme approaches the replication fork, it faces double-stranded DNA with its weak subsite. The data indicate that the different binding subsites are located sequentially, with the weak binding subsite constituting the entry site for double-stranded DNA of the replication fork.

The DnaB protein is an essential replication protein in *Escherichia coli* (1) which is involved in both the initiation and elongation stages of DNA replication (2–4). The protein is the *E. coli* primary replicative helicase, *i.e.* the factor responsible for unwinding the duplex DNA in front of the replication fork (5, 6). The DnaB protein is the only helicase required to reconstitute DNA replication *in vitro* from the chromosomal origin of replication. In the complex with ssDNA,[1] the DnaB protein forms a "mobile replication promoter." This nucleoprotein complex is specifically recognized by the primase in the initial stages of the priming reaction (1).

In solution, the native DnaB protein exists as a stable hexamer, composed of six identical subunits (7–9). Sedimentation equilibrium, sedimentation velocity, and nucleotide cofactor binding studies show that the DnaB helicase exists as a stable hexamer in a large protein concentration range, specifically stabilized by magnesium cations (7, 8). Hydrodynamic and electron microscopy data indicate that six protomers aggregate with cyclic symmetry in which the protomer-protomer contacts are limited to only two neighboring subunits (7, 10, 11). Sedimentation velocity and electron microscopy studies reveal that the DnaB hexamer undergoes dramatic conformational changes upon binding AMP-PNP and ssDNA, and provide direct evidence of the presence of long range allosteric interactions in the hexamer, encompassing all six subunits of the enzyme (8, 11).

Recently, we obtained the first estimate of the stoichiometry of the DnaB helicase-ssDNA complex and the mechanism of the binding (12–14). Using the quantitative fluorescence titration method, we determined that the DnaB helicase binds ssDNA with a stoichiometry of 20 ± 3 nucleotides/DnaB hexamer and that this stoichiometry is independent of the type of nucleic acid base (13). Our thermodynamic studies of binding of ssDNA oligomers to the DnaB hexamer show that the enzyme has a single, strong binding site for ssDNA (12). The results also show that the same binding site is used in the binding to oligomers and polymer nucleic acids (12, 13). Moreover, photocross-linking experiments indicate that the ssDNA binding site is located predominately, if not completely, on a single subunit of the hexamer (12, 13).

The reaction catalyzed by a helicase, the unwinding of a duplex DNA, must take place in the DNA binding site. The fact that the helicase uses the same single DNA binding site, when forming a complex with polymer ssDNAs, oligomers, and replication fork substrates, indicates a complex structure of the nucleic acid binding site that can accommodate both ssDNA and dsDNA.

In this communication, we report the analysis of interactions between the DnaB helicase and DNA within the total DNA binding site of the enzyme. We present direct evidence that the total DNA binding site of the helicase is structurally and functionally heterogeneous. The total binding site is built of two subsites, each encompassing approximately 10 nucleotide residues. We provide direct proof that the DnaB hexamer binds ssDNA in a strictly single orientation, with respect to the polarity of the sugar-phosphate backbone of the nucleic acid. The results indicate that the binding subsites are sequentially located along the nucleic acid lattice, with the weak binding subsite constituting an entry site for the duplex part of the replication fork.

MATERIALS AND METHODS

Reagents and Buffers—All solutions were made with distilled and deionized >18 megaohms (Milli-Q Plus) water. All chemicals were reagent grade. Buffer T2 is 50 mM Tris adjusted to pH 8.1 with HCl, 5 mM $MgCl_2$, 10% glycerol. Buffer H is 50 mM Hepes adjusted to pH 8.1 with HCl, 5 mM $MgCl_2$, 10% glycerol. The temperature, AMP-PNP, and

* This work was supported in part by National Institutes of Health Grant GM-46679 (to W. B.); John Sealy Memorial Endowment Fund Grant 2545-95; and NIEHS, National Institutes of Health, Grant 5P30ES06676. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Dept. of Human Biological Chemistry and Genetics, University of Texas Medical Branch at Galveston, 301 University Blvd., Galveston, TX 77555-1053.

[1] The abbreviations used are: ssDNA, single-stranded DNA; dsDNA, double-stranded DNA; AMP-PNP, β,γ-imidoadenosine-5′-triphosphate; CPM, 7-diethylamino-3-(4′-maleimidylphenyl)-4-methylcoumarin; Fl, fluorescein.

salt concentrations are indicated in the text. The fluorescent markers, CPM, and fluorescein 5′-isothiocyanate, used in the modification, were purchased from Molecular Probes (Eugene, OR).

DnaB Protein—The *E. coli* DnaB protein was purified, as described previously by us (7, 15–17). The concentration of the protein was spectrophotometrically determined, using extinction coefficient $\epsilon_{280} = 1.85 \times 10^5$ cm^{-1} M^{-1} (hexamer) (7).

Site-directed Mutagenesis of the DnaB Helicase—Replacement of the arginine residues at position 14 from the N terminus of the DnaB protein and obtaining the DnaB protein variant, R14C, were performed using the plasmid RLM1038, harboring the gene of the wild type DnaB helicase, generously provided by Dr. R. McMacken. The site-directed mutagenesis was accomplished in the NIEHS Center facility (National Institutes of Health) directed by Dr. T. Wood.

Labeling the DnaB R14C Variant with Fluorescent Markers—Labeling of the 6 cysteine residues of the DnaB variant, R14C hexamer, with CPM was performed in H buffer (pH 8.1, 100 mM NaCl, 5 mM $MgCl_2$, 10% glycerol) at 4 °C. The fluorescent label was added from the stock solution to the molar ratio of the CPM/R14C ~25. The mixture was incubated for 4 h, with gentle mixing. After incubation, the protein was precipitated with ammonium sulfate and dialyzed overnight against buffer T2. Any remaining free dye was removed from the modified R14C-CPM by applying the sample on a DEAE-cellulose column and eluting with buffer T2 containing 500 mM NaCl. The degree of labeling was determined by absorbance of the marker at 394 nm using the extinction coefficient of CPM, $\epsilon_{394} = 27 \times 10^3$ cm^{-1} M^{-1}, providing the value of 5.8 ± 0.1 of CPM per DnaB hexamer.[2]

Nucleic Acids—All nucleic acids were purchased from Midland Certified Reagents (Midland, TX). The etheno-derivatives of nucleic acids were obtained by modification with chloroacetaldehyde (12, 18). Oligomer dT(pT)$_{19}$, labeled at the 5′-end with fluorescein, 5′-Fl-dT(pT)$_{19}$, was synthesized using fluorescein phosphoramidate (Glen Research). Labeling of the 3′-end was performed by synthesizing dT(pT)$_{19}$ with the last residue at the 3′-end of the oligomer having the amino group on a six-carbon linker. The amino group was subsequently modified with fluorescein 5′-isothiocyanate to obtain dT(pT)$_{19}$-Fl-3′. The degree of labeling was determined by absorbance at 494 nm (pH 9), using the extinction coefficient, 7.6×10^4 M^{-1} cm^{-1} (13). The same procedures were used for labeling the 5′- and 3′-ends of the dA(pA)$_9$. The concentrations of labeled oligomers were spectrophotometrically determined at 260 nm (pH 8.1), using extinction coefficients, 1.76×10^5 M^{-1} cm^{-1} and 11.4×10^5 M^{-1} cm^{-1}, respectively (13). The concentrations of dεA(pεA)$_9$, dεA(pεA)$_8$, dεA(pεA)$_7$, dεA(pεA)$_6$, dεA(pεA)$_5$, dεA(pεA)$_4$, and dεA(pεA)$_3$ were determined using extinction coefficients 37×10^3, 33.3×10^3, 29.6×10^3, 25.9×10^3, 22.2×10^3, 18.5×10^3, and 14.8×10^3 M^{-1} cm^{-1} at 257 nm, respectively (12, 13, 19). Labeling the 5′-ends of ssDNA oligomers with ^{32}P was performed using the standard procedure (12).

Sedimentation Velocity Measurements—Analytical sedimentation experiments were performed using an Optima XL-A analytical ultracentrifuge. Analyses of the sedimentation runs were performed as we previously described (8, 9, 13). The reported values of sedimentation coefficients were corrected to standard conditions, $s_{20,w}$, for solvent density and viscosity (7).

Fluorescence Measurements—All steady-state fluorescence measurements were performed using the SLM-Aminco 48000S and 8100 spectrofluorometers (20). The emission spectra were corrected for a wavelength dependence of the instrument response using a software provided by the manufacturer. The binding of the DnaB protein was followed by monitoring the fluorescence of the etheno-derivatives of ssDNA oligomers ($\lambda_{ex} = 325$ nm, $\lambda_{em} = 410$ nm). All titration points were corrected for dilution and, if necessary, for inner filter effect using the formula (15),

$$F_{i_{cor}} = (F_i - B_i)\left(\frac{V_i}{V_o}\right)10^{0.5b(A_{i_{ex}})} \qquad \text{(Eq. 1)}$$

where $F_{i_{cor}}$ is the corrected value of the fluorescence intensity at a given point of titration i, F_i is the experimentally measured fluorescence intensity, B_i is the background, V_i is the volume of the sample at a given titration point, V_o is the initial volume of the sample, b is the total length of the optical path in the cuvette expressed in centimeters, and $A_{i_{ex}}$ is the absorbance of the sample at the excitation wavelength. Computer fits were performed using KaleidaGraph software (Synergy Software, PA) and Mathematica (Wolfram Research, IL). The relative fluorescence increase of the nucleic acid, ΔF, upon binding the DnaB protein is defined by the equation,

$$\Delta F = \frac{(F_{i_{cor}} - F_o)}{F_o} \qquad \text{(Eq. 2)}$$

where $F_{i_{cor}}$ is defined by Equation 1, and F_o is the initial value of the fluorescence of the same solution.

All steady-state fluorescence anisotropy measurements were performed in the L format, using Glan-Thompson polarizers placed in the excitation and emission channels. The fluorescence anisotropy, r, of the sample was calculated by the equation,

$$r = \frac{(I_{VV} - GI_{VH})}{(I_{VV} + 2GI_{VH})} \qquad \text{(Eq. 3)}$$

where I is the fluorescence intensity, and the first and second subscripts refer to vertical (V) polarization of the excitation and vertical (V) or horizontal (H) polarization of the emitted light (16). The factor $G = I_{HV}/I_{HH}$ corrects for the different sensitivity of the emission monochromator for vertically and horizontally polarized light (21). The limiting fluorescence anisotropies of fluorophores, r_o, were determined by measuring the anisotropy of a given sample at different solution viscosity, adjusted by sucrose or glycerol, and extrapolating to viscosity = ∞, using the Perrin equation (22).

Determination of the Average Fluorescence Energy Transfer Efficiency from CPM on the Small 12-kDa Domains of the DnaB Hexamer to the Fluorescein Residue Attached at the 5′- or 3′-End of the ssDNA Oligomers—The efficiency of the fluorescence radiationless energy transfer, E, from CPM (donor), located on the small 12-kDa domains of the DnaB protein variant R14C, to the fluorescein (acceptor), located at the 5′- or 3′-end of dT(pT)$_{19}$, bound in the DNA binding site of the helicase, has been determined using two independent methods. The fluorescence of the donor in the presence of the acceptor, F_{DA}, is related to the fluorescence of the same donor, F_D, in the absence of the acceptor by the equation,

$$F_{DA} = (1 - \nu_D)F_D + F_D\nu_D(1 - E_D) \qquad \text{(Eq. 4)}$$

where ν_D is the fraction of donors in the complex with the acceptor, and E_D is the average fluorescence energy transfer from donor to acceptor, determined from the quenching of the donor fluorescence. Thus, the average transfer efficiency, E_D, obtained from the quenching of the CPM fluorescence upon binding of the labeled ssDNA oligomer, is obtained by rearranging Equation 4,

$$E_D = \left(\frac{1}{\nu_D}\right)\left(\frac{F_D - F_{DA}}{F_D}\right) \qquad \text{(Eq. 5)}$$

where, in the considered case, F_D and F_{DA} are the values of the CPM fluorescence intensity in the absence and presence of bound 5′-Fl-dT(pT)$_{19}$ or dT(pT)$_{19}$-Fl-3′. The value of ν_D has been determined using the binding constants of the 20- and 10-mers for the DnaB helicase measured in the same solution conditions (13).

In the second independent method, the average fluorescence transfer efficiency, E_A, has been determined, using a sensitized acceptor fluorescence by measuring the fluorescence intensity of the acceptor (fluorescein) excited at 435 nm, where the donor (CPM) predominantly absorbs, in the absence and presence of R14C-CPM. The fluorescence intensities of the acceptor in the absence, F_A, and presence, F_{AD}, of the donor are defined as follows,

$$F_A = I_o\epsilon_A C_{AT}\phi_F^A \qquad \text{(Eq. 6)}$$

and

$$F_{AD} = (1 - \nu_A)F_A + I_o\epsilon_A\nu_A C_{AT}\phi_B^A + I_o\epsilon_D C_{DT}\nu_D\phi_B^A E_A \qquad \text{(Eq. 7)}$$

where I_o is the intensity of incident light, C_{AT} and C_{DT} are the total concentrations of acceptor and donor, ν_A is the fraction of acceptors in the complex with donors, ϵ_A and ϵ_D are the molar absorption coefficients of acceptor and donor at the excitation wavelength (435 nm), respectively; ϕ_F^A and ϕ_B^A are the quantum yields of the free and bound acceptor; and E_A is the average transfer efficiency determined by acceptor-sensitized emission. All quantities in Equations 6 and 7 can be experimentally determined. For the case considered in this work, the acceptor is practically completely saturated with the donor, *i.e.* $\nu_A = 1$. Thus, for $\nu_A = 1$, dividing Equation 7 by Equation 6 and rearranging provides the average transfer efficiency as described by the following.

[2] S. Rajendran, M. J. Jezewska, and W. Bujalowski, manuscript in preparation.

$$E_A = \left(\frac{1}{\nu_D}\right)\left(\frac{\epsilon_A C_{AT}}{\epsilon_D C_{DT}}\right)\left[\left(\frac{\phi_F^A}{\phi_B^A}\right)\left(\frac{F_{AD}}{F_A}\right) - 1\right] \quad \text{(Eq. 8)}$$

It should be pointed out that the energy transfer efficiencies, E_D and E_A, are apparent quantities. E_D is a fraction of the photons absent in the donor emission as a result of the presence of an acceptor, including transfer to the acceptor and possible nondipolar quenching processes induced by the presence of the acceptor, and E_A is a fraction of all photons absorbed by the donor that were transferred to the acceptor. The true Förster energy transfer efficiency, E, is a fraction of photons absorbed by the donor and transferred to the acceptor in the absence of any additional nondipolar quenching resulting from the presence of the acceptor (22). The value of E is related to the apparent quantities of E_D and E_A, by the following (23).

$$E = \frac{E_A}{(1 - E_D + E_A)} \quad \text{(Eq. 9)}$$

Thus, measurements of the transfer efficiency, using both methods, are not alternatives but parts of the analysis used to obtain the true efficiency of the fluorescence energy transfer process, E.

The fluorescence energy transfer efficiency between donor and acceptor dipoles is related to the distance, R, separating the dipoles by the equation,

$$E = \frac{R_o^6}{R_o^6 + R^6} \quad \text{(Eq. 10)}$$

where $R_o = 9790\ (\kappa^2\, n^{-4}\, \phi_d\, J)^{1/6}$ is the so called Förster critical distance (in angstroms), the distance at which the transfer efficiency is 50%; κ^2 is the orientation factor; ϕ_d is the donor quantum yield in the absence of the acceptor; and n is the refractive index of the medium (n = 1.4) (22). The overlap integral, J, characterizes the resonance between the donor and acceptor dipoles.

The fluorescence transfer efficiency of chemically identical donor and acceptor pairs, characterized by the same quantum yields, depends on the distance between the donor and acceptor, R, and the factor, κ^2, describing the mutual orientation of the donor and acceptor dipoles (22). Although in the work presented in this paper we are interested in relative distances between donors and acceptors, evaluation of κ^2 allowed us to estimate the effect of the orientation factor on the differences between the studied donor-acceptor distances. The factor κ^2 cannot be experimentally determined; however, the upper (κ^2_{max}) and lower (κ^2_{min}) limits of κ^2 can be obtained from the measured limiting anisotropies of the donor and acceptor and the calculated axial depolarization factors, using the procedure described by Dale *et al.* (24). When both axial depolarization factors are positive, κ^2_{max} and κ^2_{min} can be calculated from $\kappa^2_{max} = (2/3)(1 + \langle d^X_D\rangle + \langle d^X_A\rangle + 3\langle d^X_D\rangle\langle d^X_A\rangle)$ and $\kappa^2_{min} = (2/3)(1 - (1/2)(\langle d^X_D\rangle + \langle d^X_A\rangle)$, where $\langle d^X_D\rangle$ and $\langle d^X_A\rangle$ are the axial depolarization factors for the donor and acceptor, respectively (24). The axial depolarization factors have been calculated as square roots of the ratios of the limiting anisotropies of the donors (CPM on the DnaB helicase) and acceptors (fluorescein at the 5′- or 3′-end of the ssDNA oligomers) and their corresponding fundamental anisotropies (17). For two chemically identical donor-acceptor pairs, characterized by the same R_o (the same κ^2, ϕ_d, and J), the differences in the transfer efficiencies, E_1 and E_2, result exclusively from the different distances between the donor and acceptor, R_1 and R_2. The relative ratio of the two distances is then defined by using Equation 10 as follows.

$$\frac{R_1}{R_2} = \left\{\frac{[(1 - E_1)E_2]}{[(1 - E_2)E_1]}\right\}^{1/6} \quad \text{(Eq. 11)}$$

Determination of Rigorous Thermodynamic Binding Isotherms and Absolute Stoichiometries of the DnaB Helicase-ssDNA Complexes—In this work, we followed the binding of the DnaB protein to the ssDNA oligomers by monitoring the fluorescence increase, ΔF, of ssDNA etheno-derivatives upon the complex formation. Proteins and nucleic acids may form complexes characterized by different spectroscopic properties, particularly when multiple ligand binding processes are studied. In applying spectroscopic methods to monitor the ligand macromolecule interactions, one should not assume strict proportionality between the observed signal change and the degree of binding unless the existence of such proportionality has been shown (15). The general method to obtain thermodynamically rigorous estimates of the average degree of binding of the protein per ssDNA oligomer, $\Sigma\nu_i$, and the free protein concentration, P_F, has been previously described by us (8, 15, 25). Briefly, the experimentally observed ΔF has a contribution from each of the different possible "i" complexes of the DnaB hexamer with a nucleic acid. Thus, the observed fluorescence increase is functionally related to $\Sigma\nu_i$ by the equation,

$$\Delta F = \Sigma\, \nu_i \Delta F_{i_{max}} \quad \text{(Eq. 12)}$$

where $\Delta F_{i_{max}}$ is the molecular parameter characterizing the maximum fluorescence increase of the nucleic acid with the DnaB protein bound in complex i. The same value of ΔF, obtained at two different total nucleic acid concentrations, N_{T_1} and N_{T_2}, indicates the same physical state of the nucleic acid, *i.e.* the degree of binding, $\Sigma\nu_i$, and the free DnaB protein concentration, P_F, must be the same. The value of $\Sigma\nu_i$ and P_F is then related to the total protein concentrations, P_{T_1} and P_{T_2}, and the total nucleic acid concentrations, N_{T_1} and N_{T_2}, at the same value of ΔF, by the following equations,

$$\Sigma\, \nu_i = \frac{(P_{T_2} - P_{T_1})}{(N_{T_2} - N_{T_1})} \quad \text{(Eq. 13)}$$

$$P_F = P_{T_x} - (\Sigma\, \nu_i) N_{T_x} \quad \text{(Eq. 14)}$$

where x = 1 or 2 (12, 20).

RESULTS

Micrococcal Nuclease Digestion Reveals Large Structural Heterogeneity within the DNA Binding Site of the E. coli DnaB Helicase—Quantitative fluorescence titrations and photo-cross-linking experiments, using ssDNA oligomers, showed that the DnaB hexamer has a single ssDNA binding site encompassing 20 ± 3 nucleotide residues and located predominantly on a single subunit (12–14). The first evidence of the structural heterogeneity within the DNA binding site came from nuclease digestion-protection studies of the DNA in the complex with the helicase. In the first set of experiments, the complex of the DnaB hexamer with the 20-mer $dT(pT)_{19}$ labeled at its 5′-end with ^{32}P in the presence of 1 mM AMP-PNP was subjected to micrococcal nuclease digestion as a function of time. The protein was in molar excess over the 20-mer to ensure complete saturation of the nucleic acid. Fig. 1*a* shows the polyacrylamide sequencing gel of $dT(pT)_{19}$ after digestion with the nuclease, at different time intervals, in the absence and presence of the helicase. In the absence of the helicase, in our solution conditions, the 20-mer was digested within 20 min. A dramatically different behavior was observed in the presence of the enzyme. The digestion process was less efficient, indicating significant protection of the nucleic acid against the nuclease by the enzyme. Moreover, at prolonged digestion times, a nucleic acid fragment of 10 or 11 nucleotide residues was strongly protected by the helicase. At the longest times, this was the major nucleic acid fragment on the gel, resistant to further nuclease action (Fig. 1*a*).

The size of the protected fragment was not dependent upon the length or type of base of the oligomer bound to the DnaB protein, indicating that protection against the nuclease digestion is limited to the nucleic acid bound within the single DNA binding site of the helicase. Fig. 1*b* shows polyacrylamide sequencing gels of $dA(pA)_{69}$ after digestion with the nuclease, at different time intervals, and in the absence and presence of the helicase. As in the case of $dT(pT)_{19}$, the only predominant oligomer protected by the helicase in the complex with $dA(pA)_{69}$, after prolonged digestion, is a ssDNA fragment, 10 or 11 nucleotide residues long.

These data indicate that, within the total DNA binding site of the DnaB helicase, approximately half of the ~20 nucleotide residues occluded by the helicase are bound differently than the remaining half, resulting in the observed nuclease digestion pattern. Thus, these results indicate that the total DNA binding site of the DnaB helicase is built of two structurally and possibly functionally different binding subsites (see below).

Binding of 10-mer, dεA(pεA)$_9$, to the DnaB Helicase—To de-

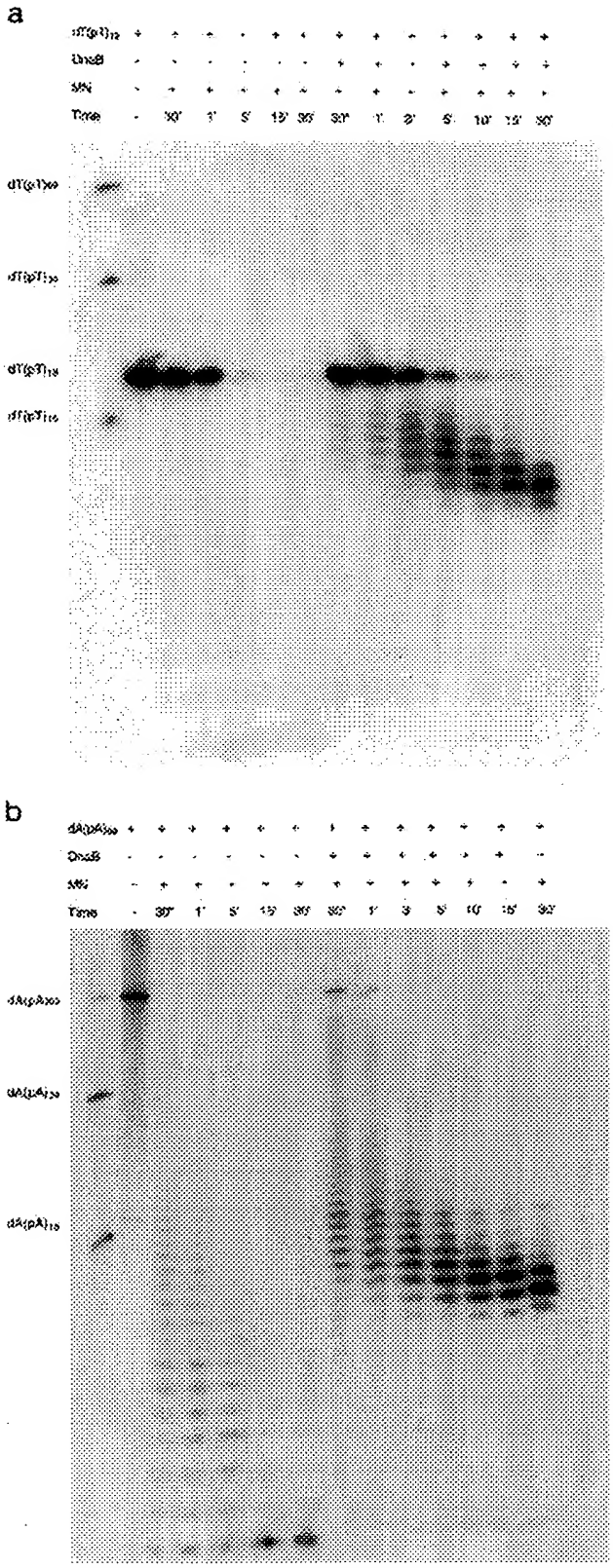


FIG. 1. *a*, autoradiogram of the 15% sequencing polyacrylamide gel electrophoresis of the 5′-[^{32}P](dT)$_{20}$ and DnaB protein-5′-[^{32}P](dT)$_{20}$ complexes, after micrococcal nuclease (*MN*) digestion in buffer T2 (pH 8.1, 4 °C) containing 100 mM NaCl, 1 mM $CaCl_2$, and 1 mM AMP-PNP. The concentration of the protein and the 20-mer are 1×10^{-6} M (hexamer) and 5×10^{-7} M (oligomer). Oligomers of different lengths are included in *lane 1* as molecular markers. *Lanes 2–6* show the different digestion times without the DnaB helicase. *Lane 2*, 5′-[^{32}P](dT)$_{20}$, 0 s; *lane 3*, 30 s; *lane 4*, 60 s; *lane 5*, 300 s; *lane 6*, 900 s; *lane 7*, 1800 s. *Lanes 8–14* show the complex 5′-[^{32}P](dT)$_{20}$-DnaB helicase at different digestion times. *Lane 8*, 30 s; *lane 9*, 60 s; *lane 10*, 180 s; *lane 11*, 300 s; *lane 12*, 600 s; *lane 13*, 900 s; *lane 14*, 1800 s. *b*, autoradiogram of the 15% sequencing polyacrylamide gel electrophoresis of the 5′-[^{32}P](dA)$_{70}$ and DnaB protein-5′-[^{32}P](dA)$_{70}$ complexes, after micrococcal nuclease digestion in buffer T2 (pH 8.1, 4 °C) containing 100 mM NaCl, 1 mM $CaCl_2$ and 1 mM AMP-PNP. The concentration of the protein and the 70-mer are 1×10^{-6} M (hexamer) and 5×10^{-7} M (oligomer), respectively. *Lanes 2–7* show the 5′-[^{32}P](dA)$_{70}$ at different digestion times without the DnaB helicase. *Lane 2*, 5′-[^{32}P](dA)$_{70}$, 0 s; *lane 3*, 30 s; *lane 4*, 60 s; *lane 5*, 300 s; *lane 6*, 900 s; *lane 7*, 1800 s. *Lanes 8–14* show the complex 5′-[^{32}P](dA)$_{70}$-DnaB helicase at different digestion times. *Lane 8*, 30 s; *lane 9*, 60 s; *lane 10*, 180 s; *lane 11*, 300 s; *lane 12*, 600 s; *lane 13*, 900 s; *lane 14*, 1800 s.

termine whether or not there is a difference in affinities between the two subsites of the total DNA binding site of the helicase that could result in their different functional roles in the enzyme activities, we performed quantitative thermodynamic studies of the binding of a ssDNA oligomer containing only 10 residues to the DnaB hexamer. These are the partial nucleic acid ligands that can only interact with half of the total binding site of the enzyme. Fluorescence titrations of dεA(pεA)$_9$ with the DnaB helicase, at five different nucleic acid concentrations, in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP, are shown in Fig. 2*a*. As the nucleic acid concentration increases, the same relative fluorescence increase is reached at higher DnaB protein concentrations. The selected nucleic acid concentrations provide the separation of binding isotherms up to a relative fluorescence increase of ~4.1, with the plateau at the maximum relative fluorescence increase, $\Delta F_{max} = 4.3 \pm 0.2$. To obtain thermodynamically rigorous binding parameters, independent of any assumption about the relationship between the observed signal and the degree of binding, $\Sigma \nu_i$, the fluorescence titration curves shown in Fig. 2*a* were analyzed, using the approach outlined under "Materials and Methods." Fig. 2*b* shows the dependence of the observed relative fluorescence increase as a function of the average number of the DnaB hexamers bound per oligomer. The plot is linear, indicating that, in the studied binding density range, there is a very similar enhancement of the nucleic acid fluorescence upon the binding of the oligomer to the DnaB protein. The value of $\Sigma \nu_i$ could be determined up to ~90% of the observed signal change. Short extrapolation to the maximum value of the fluorescence increase provides the stoichiometry of the complex. Thus, the data show that only one 10-mer strongly binds to the DnaB hexamer, indicating that the structural differences between the subsites, resulting in protection of ~10 nucleotide residues of ssDNA in the total DNA binding site against nuclease digestion, are reflected in the large differences in the affinities between the two DNA binding subsites. The *solid lines* in Fig. 2*a* are computer fits of the binding isotherms to a single binding site that provide the binding constant $K = (1.7 \pm 0.3) \times 10^7$ M^{-1}. Comparison with the binding constant for the 20-mer, dεA(pεA)$_{19}$, previously obtained by us in the same solution conditions ($K = 3 \times 10^7$ M^{-1}), shows that the 10-mer binds with an affinity very similar to the 20-mer (13). Thus, the data indicate that the predominant part of the free energy of binding the DnaB helicase to ssDNA comes from the interactions of the nucleic acid with the strong binding subsite of the enzyme (see "Discussion").

Quantitative fluorescence titrations at a very high concentration of dεA(pεA)$_9$ (~8×10^{-6} M (oligomer)) did not show detectable binding of the additional 10-mer (Fig. 2*a*). Titrations at higher nucleic acid concentrations are very difficult because they require very high concentrations of stock solutions of the DnaB protein, which are beyond the attainable solubility of the

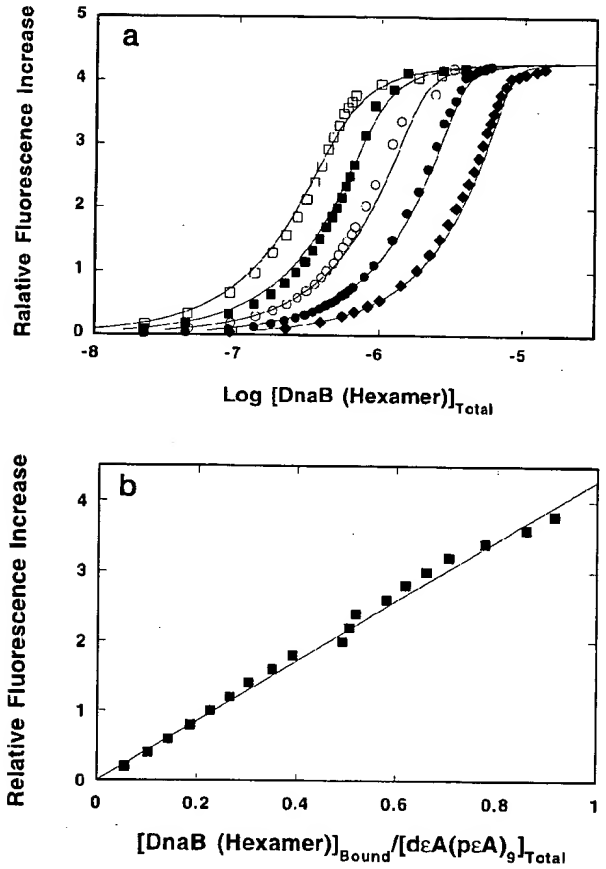


FIG. 2. *a*, fluorescence titrations of dεA(pεA)$_9$ with the DnaB protein monitored by the increase of the nucleic acid fluorescence in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP, at three different nucleic acid concentrations (oligomer): 4.5×10^{-7} M (□); 9.0×10^{-7} M (■); 1.9×10^{-6} M (○); 3.6×10^{-6} M (●); 8.0×10^{-6} M (◆). *Solid lines* are computer fits of the single-site binding isotherm, $\Delta F = \Delta F_{max}(KP_F/(1 + KP_F))$, with intrinsic binding constant $K_1 = 1.7 \times 10^7$ M^{-1} and $\Delta F_{max} = 4.3$. *b*, the dependence of the relative increase of the dεA(pεA)$_9$ fluorescence upon the average number of DnaB helicase hexamers bound per oligomer (■). The absolute value of the average number of DnaB helicase hexamers bound per oligomer, $\Sigma\nu_i$, has been determined using the thermodynamically rigorous approach described under "Materials and Methods." The *solid line* is a computer fit using the single-site binding isotherm ($\Delta F = \Delta F_{max}(K_1P_F/(1 + K_1P_F))$; $\Sigma\nu_i = K_1P_F/(1 + K_1P_F)$) with $K_1 = 1.7 \times 10^7$ M^{-1} and $\Delta F_{max} = 4.3$; P_F is the free dεA(pεA)$_9$ concentration.

protein. Therefore, we used the analytical centrifugation technique to assess the affinity of DNA to the second subsite. In these experiments, we used a 10-mer, dA(pA)$_9$, labeled at the 5′- or 3′-end with fluorescein (see "Materials and Methods"). This approach allowed us to monitor exclusively the nucleic acid and the protein-nucleic acid complex without the interference of the protein and AMP-PNP absorbance. The sedimentation velocity profiles (monitored at 515 nm) of the DnaB helicase-dA(pA)$_9$-3′-Fl mixture at the nucleic acid and helicase concentrations of 9×10^{-5} and 2×10^{-5} M, respectively, in buffer T2 (pH 8.1, 20 °C) containing 100 mM NaCl and 1 mM AMP-PNP, are shown in Fig. 3. The sedimentation run was performed at 60,000 rpm. It is clear that, initially, two independently moving boundaries exist. The slow moving boundary has a sedimentation coefficient of $s_{20,w} = 1.4 \pm 0.2$, which is the $s_{20,w}$ value of the free dA(pA)$_9$-3′-Fl. The fast moving boundary contains dA(pA)$_9$-3′-Fl in the complex with the DnaB helicase. We have previously shown that the DnaB hexamer fully preserves its hexameric structure in the complex with the ssDNA

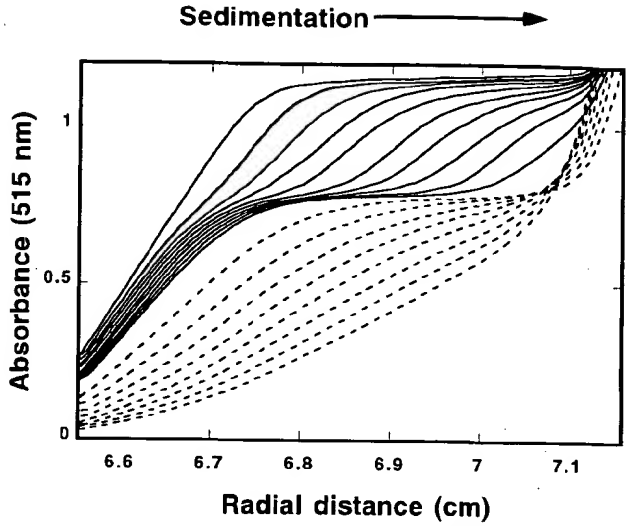


FIG. 3. **Absorption profiles at 515 nm of the sedimentation velocity runs of the dA(pA)$_9$-3′-Fl-DnaB protein complex in buffer T2 (pH 8.1, 20 °C) containing 100 mM NaCl and 1 mM AMP-PNP, at a 4.5:1 molar excess of the nucleic acid over the enzyme.** The concentration of the DnaB hexamer is 2×10^{-5} M (hexamer), and the concentration of the dA(pA)$_9$-3′-Fl is 9×10^{-5} M (oligomer). *Solid lines* are initial scans of the samples, which include slow and fast moving boundaries of the nucleic acid and the complex, respectively. *Dashed lines* are scans of the sample after the fast moving boundary reached the bottom of the cell. The initial part of the last scan indicates the location of the base line (time interval was 8 min; 60,000 rpm).

(8, 13). After the fast moving boundary reaches the cell bottom, only the slow moving boundary of the free dA(pA)$_9$-3′-Fl still remains (*dashed lines*). Notice that during the sedimentation process, the boundary of the complex migrates in the field of the constant free 10-mer concentration $[T]_{Free} \gg 1/K_1$, thus assuring that the enzyme always has the strong binding subsite saturated with the nucleic acid. At 515 nm, one monitors exclusively the total concentration of the 10-mer. Comparison of the contributions of the slow and fast moving boundaries with the total absorption of the sample shows that 36% of the total nucleic acid concentration migrated in the fast moving boundary (Fig. 3). From the known total concentration of the DnaB helicase in the sample, the stoichiometry of the complex is calculated to be 1.6 ± 0.2, which indicates that at this 10-mer concentration we observed significant saturation (60%) of the second DNA binding subsite.

Because we know the free nucleic acid concentration from the absorbance of the slowly moving boundary, we can estimate the macroscopic ssDNA binding constant for the second subsite. In the considered case, the first binding subsite of the helicase is completely saturated with the 10-mer. Therefore, the partition function of the system, Z, and the degree of binding to the second subsite, ν_2, are as follows,

$$Z = K_1[T]_{Free} + K_1K_2[T]_{Free}^2 \quad \text{(Eq. 15)}$$

$$\nu_2 = \frac{K_1K_2[T]_{Free}^2}{Z} \quad \text{(Eq. 16)}$$

and K_2 is defined as follows.

$$K_2 = \frac{\nu_2}{\{[T]_{Free}(1 - \nu_2)\}} \quad \text{(Eq. 17)}$$

Introducing the values of $\nu_2 = 0.6$ and $[T]_{Free} = 5.8 \times 10^{-5}$ M, obtained from the sedimentation velocity experiments, provides the value of $K_2 = (2.6 \pm 1) \times 10^4$ M^{-1}. A similar value of the binding constant K_2 of the dA(pA)$_9$-3′-Fl and 5′-Fl-dA(pA)$_9$

for the weak binding subsite has been obtained using lower and higher concentrations of the nucleic acids. Thus, the data show that the affinity of ssDNA for the second subsite is ~3 orders of magnitude lower than the affinity for the strong binding subsite.

Interactions of ssDNA Oligomers Having Different Lengths with the Strong ssDNA Binding Subsite—To obtain further insight into the interactions of the DNA in the strong binding subsite, we performed quantitative fluorescence titrations of a series of ssDNA oligomers of different lengths. Fluorescence titrations of dεA(pεA)$_8$, dεA(pεA)$_7$, dεA(pεA)$_6$, dεA(pεA)$_5$, dεA(pεA)$_4$, and dεA(pεA)$_3$, with the DnaB helicase in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP, are shown in Fig. 4*a*. For comparison, the fluorescence titration of the 10-mer, dεA(pεA)$_9$, with DnaB is also included. With the decreasing number of residues, the relative maximum fluorescence changes, and the affinity decreases. In the case of 9- and 8-mers, the maximum fluorescence change, ΔF_{max}, upon saturation with the helicase, is still similar to the one determined for the 10-mer. However, the affinity is lower than the affinity of the 10-mer and is characterized by binding constants $K_9 = (1.5 \pm 0.5) \times 10^7$ M^{-1} and $K_8 = (8 \pm 2) \times 10^6$ M^{-1}, respectively. A dramatic drop in the affinity and maximum relative fluorescence increase is observed in the case of the 7- and 6-mer (Fig. 4*a*; see Table I). No detectable binding to the helicase occurs in the case of 5- and 4-mers (Fig. 4*a*). Titrations at very high concentrations of the 5- and 4-mer could only provide a semi-quantitative estimate of the affinities, due to the required DnaB concentration beyond the solubility of the protein; however, these experiments indicate that the binding constants for the 5- and 4-mers are not higher than 1×10^4 M^{-1} (data not shown). Fig. 4*b* shows the dependence of the natural logarithm of binding constants of studied oligomers to the DnaB protein as a function of the number of nucleotide residues in the ssDNA oligomer. The plot is nonlinear, a clear indication that the affinity is not a simple function of the length of the nucleic acid. The difference of the 2 residues between 10-mer and 8-mer causes only an ~0.3 kcal/mol decrease of the free energy of interactions. The difference in the 2 residues between 7-mer and 5-mer decreases the free energy of binding by at least ~3 kcal/mol, practically eliminating the binding of the 5-mer to the enzyme in studied solution conditions. These data show that, to efficiently bind to the strong DNA binding subsite, the nucleic acid must span 6 or 7 residues. Thus, the results indicate a complex structure of the ssDNA strong binding subsite where the direct contacts between the helicase and the nucleic acid, decisive in complex formation, are separated by 6 or 7 nucleotides (see "Discussion").

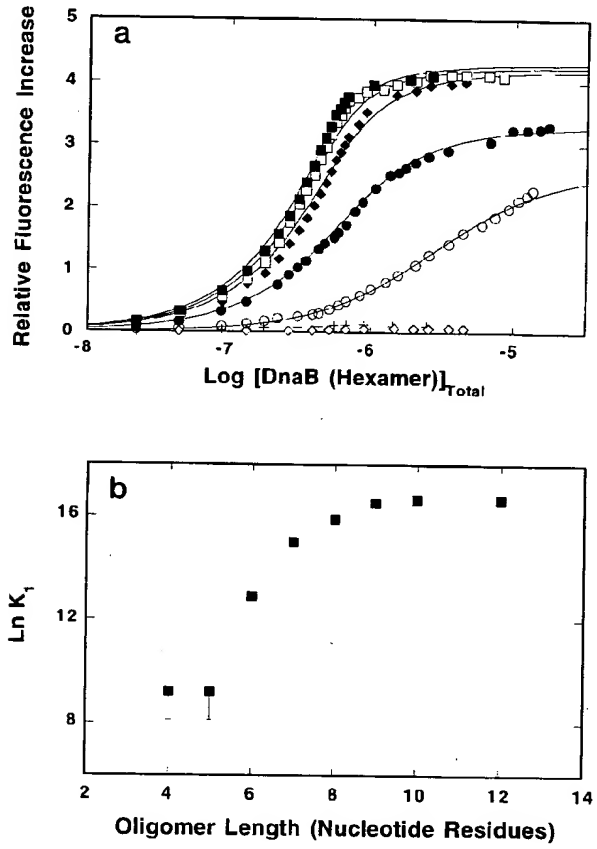


FIG. 4. *a*, fluorescence titrations of dεA(pεA)$_8$, dεA(pεA)$_7$, dεA(pεA)$_6$, dεA(pεA)$_5$, dεA(pεA)$_4$, and dεA(pεA)$_3$, with the DnaB protein monitored by the increase of the nucleic acid fluorescence in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP. Concentrations of all oligomers are 4.5×10^{-7} M (oligomer). □, dεA(pεA)$_8$; ◆, dεA(pεA)$_7$; ●, dεA(pεA)$_6$; ○, dεA(pεA)$_5$; +, dεA(pεA)$_4$; ◇, dεA(pεA)$_3$. For comparison, the fluorescence titration of dεA(pεA)$_9$ is also included (■). *Solid lines* are computer fits of the single-site binding isotherm, $\Delta F = \Delta F_{max}(K_1P_F/(1 + K_1P_F))$, with binding constants K_1 and ΔF_{max} as follows: 1.5×10^7 M^{-1} and 4.3; 8×10^6 M^{-1} and 4.3; 3.3×10^6 M^{-1} and 3.3; and 4×10^5 M^{-1} and 2.6 for the 9-, 8-, 7-, and 6-mer, respectively. *b*, the dependence of the natural logarithm of binding constant as a function of the length of the oligomer bound to the strong subsite of the DnaB helicase (■). The binding constants for 5- and 4-mers have an assigned maximum value at 1×10^4 M^{-1}, which is below the minimum affinity detectable in our fluorescence titrations (~5×10^4 M^{-1}), although the affinities of these oligomers could be characterized by even lower binding constants, as indicated by the *error bars*.

Salt Effect on the Affinity of the DnaB Helicase to ssDNA Oligomers—Fluorescence titrations of dεA(pεA)$_8$ with the DnaB helicase in buffer T2 (pH 8.1, 10 °C), containing 1 mM AMP-PNP and different NaCl concentrations, are shown in Fig. 5*a*. As the salt concentration increases, the isotherms shift toward higher total DnaB protein concentrations, indicating a decreasing affinity of the protein-nucleic acid complex at higher salt concentrations. It should also be noted that ΔF_{max} is lower at higher salt concentrations, decreasing from 4.3 ± 0.2 at 100 mM to 2.8 ± 0.2 at 407 mM [NaCl]. A similar decrease of the maximum fluorescence increase upon the helicase binding has been observed for all other oligomers (data not shown).

The dependence of the logarithm of the intrinsic binding constants for 10-, 9-, 8-, 7-, and 6-mers upon the logarithm of [NaCl] (log-log plot) is shown in Fig. 5*b*. Within experimental accuracy, the plots are linear in the studied salt concentration ranges, which is different from the nonlinear behavior of the

TABLE I
Thermodynamic and fluorescence parameters of ssDNA oligomer binding to the DnaB helicase in buffer T2 (pH 8.1, 100 mM NaCl, 1 mM AMP-PNP, 10 °C; λ_{ex} = 325 nm, λ_{em} = 410 nm)

	dεA(pεA)$_9$	dεA(pεA)$_8$	dεA(pεA)$_7$	dεA(pεA)$_6$	dεA(pεA)$_5$
Stoichiometry (n)	1	1	1	1	1
Binding constant K (M^{-1})	$(1.7 \pm 0.3) \times 10^7$	$(1.5 \pm 0.5) \times 10^7$	$(8 \pm 3) \times 10^6$	$(3.3 \pm 1) \times 10^6$	$(4 \pm 1) \times 10^5$
ΔF_{max}	4.3 ± 0.2	4.3 ± 0.2	4.2 ± 0.2	3.3 ± 0.2	2.6 ± 0.4
$\frac{\partial \log K}{\partial \log[NaCl]}$	−1.4 ± 0.3	−1.5 ± 0.3	−1.4 ± 0.3	−1.5 ± 0.3	−1.4 ± 0.3

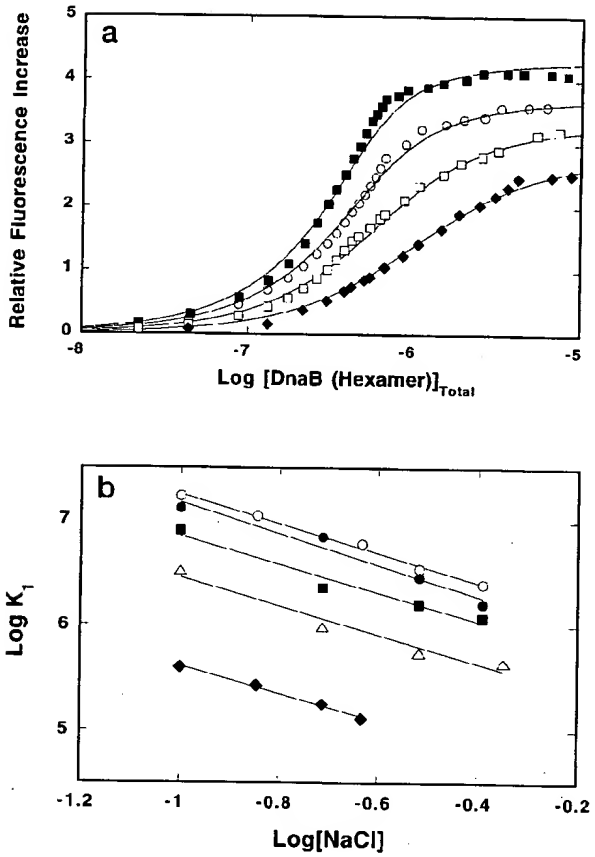


FIG. 5. *a*, fluorescence titrations of dεA(pεA)$_9$ with the DnaB protein in buffer T2 (pH 8.1, 10 °C) containing 1 mM AMP-PNP, at different NaCl concentrations as follows. ■, 100 mM; ○, 194 mM; □, 304 mM; ◆, 407 mM. *Solid lines* are computer fits using single-site binding isotherm, $\Delta F = \Delta F_{max}(K_1P_F/(1 + K_1P_F))$, with ΔF_{max} and K_1 as follows. ■, 4.3 and 1.5×10^7 M^{-1}; ○, 3.7 and 7×10^6 M^{-1}; □, 3.3 and 3×10^6 M^{-1}; ◆, 2.8 and 1.3×10^6 M^{-1}. *b*, the dependence of the intrinsic binding constant K_1 for the binding of ssDNA oligomers of different lengths to the strong binding subsite of DnaB helicase upon NaCl concentrations in solution (log-log plots) in buffer T2 (pH 8.1, 10 °C) containing 1 mM AMP-PNP. ○, dεA(pεA)$_9$; ●, dεA(pεA)$_8$; ■, dεA(pεA)$_7$; △, dεA(pεA)$_6$; ◆, dεA(pεA)$_5$.

log-log plot previously determined in the case of the 20-mer, dεA(pεA)$_{19}$ (12, 13). With increasing salt concentrations, the affinities of all oligomers decrease, indicating that the binding process is accompanied by a net release of ions with the slopes $\partial \log K/\partial \log[NaCl] = -1.4 \pm 0.4$, -1.5 ± 0.3, -1.4 ± 0.4, -1.5 ± 0.4, and -1.4 ± 0.4 for 10-, 9-, 8-, 7-, and 6-mer, respectively (27) (Table I). Thus, these data indicate that a similar number of ~1.5 ions is released upon the complex formation with each of the oligomers being long enough to provide all essential contacts with the enzyme in the binding subsite.

Previously, we determined that binding of a 20-mer, dεA(pεA)$_{19}$, which spans the entire total DNA binding site, to the DnaB helicase is accompanied by the maximum release of ~3.7 ions (13). This number is significantly higher than the ~1.4 obtained for the 10-mer (Table I). This comparison suggests that the interactions of ssDNA with the weak binding subsite are accompanied by a net release of ~two ions. Another possibility is that interactions between the strong and weak subsites, simultaneously saturated with nucleic acid in the complex with dεA(pεA)$_{19}$, result in the net release of ~two additional ions. At present, we cannot exclude either of these possibilities.

Determination of the Orientation of the E. coli DnaB Helicase with Respect to the Polarity of the Sugar-Phosphate Backbone of ssDNA, Using the Fluorescence Energy Transfer Method—Determination of the mutual orientations of proteins and nucleic acids in the complex should be based on a method that is sensitive to the differences in distances between different, specific regions of both macromolecules (17, 22). Fluorescence energy transfer between a donor and an acceptor, placed in specific locations on a protein and a nucleic acid, provides a very sensitive technique to assess the relative proximities between different regions of both macromolecules in the complex. The orientation of the DnaB helicase, in the complex with ssDNA, was determined by using the 20-mer, dT(pT)$_{19}$, labeled with fluorescein (acceptor) at its 5′- or 3′-end, respectively, and the DnaB protein variant, R14C, specifically labeled with a coumarin derivative (donor), CPM, at the small 12-kDa domain of the enzyme (see "Materials and Methods"). If the DnaB helicase binds predominantly in a single orientation, with respect to the polarity of the sugar-phosphate backbone of ssDNA, then different responses of the donor and acceptor fluorescence should be observed, depending on the different location of the acceptor on the nucleic acid.

The elongated DnaB protein monomer is built of two structural domains, small 12-kDa and large 33-kDa domains connected at the "hinge" region (28) as visualized by electron microscopy data (10, 11). In the hexamer, all protomers are oriented with their small 12-kDa domains in the same direction (10, 11). Because the protein does not have natural cysteines, we replaced arginine at position 14 from the N terminus of the protein located in the small 12-kDa domain of the enzyme with a single cysteine residue, using site-directed mutagenesis. Subsequently, this cysteine residue was specifically modified with CPM to provide R14C-CPM (see "Materials and Methods"). The selection of the modification site was directed by the fact that removal of the entire 14-amino acid fragment from the N terminus of the protein did not affect, to any extent, the biological functions of the protein (28). As a result of modification, the R14C DnaB hexamer has six CPM molecules located in the small domain of each protomer (R14C-CPM). Thus, 6 CPM residues form a ring at one end of the DnaB hexamer.

The emission spectrum of R14C-CPM strongly overlaps the absorption spectrum of the fluorescein. These spectroscopic properties of CPM make the marker an excellent fluorescence donor for fluorescein (29). The presence of unlabeled dT(pT)$_{19}$ causes very little change in the fluorescence emission spectra of R14C-CPM (λ_{ex} = 435 nm); however, the presence of R14C-CPM causes an ~2-fold decrease of the fluorescence intensity of 5′-Fl-dT(pT)$_{19}$ (λ_{ex} = 485 nm), although with the excitation at 485 nm only fluorescein on the 5′-Fl-dT(pT)$_{19}$ absorbs light (data not shown). Saturation of the 20-mer with the unlabeled DnaB protein causes only an ~8% decrease of the 5′-Fl-dT(pT)$_{19}$ fluorescence (data not shown). It is evident that, even in the absence of the energy transfer process, the presence of 6 hydrophobic CPM residues affects the quantum yield of fluorescein at the 5′-end of the ssDNA, which already suggests close proximity between the CPMs and fluorescein. The quantum yield of fluorescein is independent of the excitation wavelength between 400 and 500 nm (22). Thus, as expected, the ratio of quantum yields of 5′-Fl-dT(pT)$_{19}$ in the complex with R14C-CPM and free in solution, ϕ_B^A/ϕ_F^A, is constant and equal to 0.51 over a tested range of excitation wavelengths between 465 and 500 nm. In this spectral range of excitation, no detectable fluorescence energy transfer from CPM residues to fluorescein occurs. Thus, this ratio of quantum yields, independent of excitation wavelength, reflects the change of the emission intensity of 5′-Fl-dT(pT)$_{19}$, resulting exclusively from the formation of the complex with R14C-CPM, in the absence of the

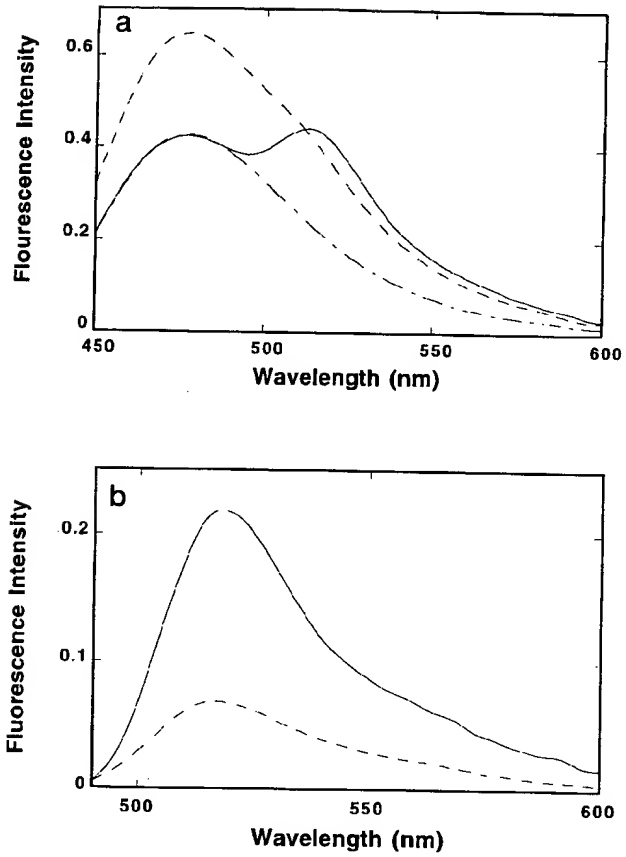


FIG. 6. *a*, sum of the fluorescence emission spectra (– – –) of DnaB R14C-CPM in the presence of unlabeled $dT(pT)_{19}$ (4.5×10^{-7} M (oligomer)) and 5′-Fl-$dT(pT)_{19}$ in the presence of R14C-CPM (without energy transfer) (λ_{ex} = 435 nm) in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP and the fluorescence emission spectrum of the complex of R14C-CPM with 5′-Fl-$dT(pT)_{19}$ (λ_{ex} = 435 nm) (——) in the same buffer. Concentrations of 5′-Fl-$dT(pT)_{19}$ and the protein were 4.5×10^{-7} M (oligomer) and 9.6×10^{-7} M (hexamer), respectively. The fluorescence emission spectrum of R14C-CPM normalized at 476 nm (peak) to the emission spectrum of the protein in the complex with 5′-Fl-$dT(pT)_{19}$ (– · – · –) is also included. *b*, sensitized emission spectrum of 5′-Fl-$dT(pT)_{19}$ (λ_{ex} = 435 nm) in the complex with R14C-CPM (——), obtained after subtraction of the normalized spectrum of R14C-CPM (see Fig. 7*a*) in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP superimposed on the fluorescence emission spectrum of 5′-Fl-$dT(pT)_{19}$ in the presence of R14C-CPM (without energy transfer) (– – –) obtained at the same excitation wavelength by multiplying the spectrum of free, labeled 20-mer by the quantum yield ratio, ϕ_D^A/ϕ_F^A = 0.51. Concentrations of 5′-Fl-$dT(pT)_{19}$ and R14C-CPM are 4.5×10^{-7} M (oligomer) and 9.6×10^{-7} M (hexamer), respectively.

energy transfer process, and can be used to obtain the spectrum of 5′-Fl-$dT(pT)_{19}$, in the presence of R14C-CPM, without the changes induced by the energy transfer process at any excitation wavelength (Equations 7 and 8).

The *dashed line* in Fig. 6*a* is the sum of the emission spectra (λ_{ex} = 435 nm) of the R14C-CPM and 5′-Fl-$dT(pT)_{19}$ in the presence of unlabeled nucleic acid and R14C-CPM (without energy transfer), respectively, in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP. The *solid line* is the fluorescence emission spectrum of the complex of R14C-CPM with 5′-Fl-$dT(pT)_{19}$ at the same concentrations of the protein and nucleic acid as in the case of the sum of independent components of the complex. Clearly, there is a dramatic difference between the sum of the independent donor and acceptor spectra and the spectrum where both donor and acceptor are placed in the same complex. The emission intensity of R14C-CPM at 476 nm in the complex with 5′-Fl-$dT(pT)_{19}$ is decreased by ~ 35%, as compared with the R14C-CPM complexed with unlabeled $dT(pT)_{19}$. The decrease of emission at 476 nm, where there is no contribution from fluorescein emission, indicates significant fluorescence energy transfer from the CPM residues located on the small 12-kDa domains of the DnaB hexamer to the fluorescein moiety placed at the 5′-end of the bound 5′-Fl-$dT(pT)_{19}$.

Comparison between the sum of the spectra of independent components of the complex and the spectrum of the complex in Fig. 6*a* shows that the fluorescence intensity of the fluorescein residue of 5′-Fl-$dT(pT)_{19}$, with the peak at 520 nm, is strongly increased in the complex with R14C-CPM (λ_{ex} = 435 nm). Recalling that fluorescein does not contribute to the CPM emission band at 476 nm, we can normalize the spectra of R14C-CPM-unlabeled $dT(pT)_{19}$ and R14C-CPM-5′-Fl-$dT(pT)_{19}$ complex at 476 nm. The difference between the normalized spectrum of R14C-CPM-unlabeled $dT(pT)_{19}$ and the spectrum of the complex R14C-CPM-5′-Fl-$dT(pT)_{19}$ provides the sensitized emission spectrum of the 5′-Fl-$dT(pT)_{19}$ bound to R14C-CPM. The emission spectrum of 5′-Fl-$dT(pT)_{19}$ in the complex with R14C-CPM, without energy transfer, with the sensitized emission spectrum of 5′-Fl-$dT(pT)_{19}$ is shown in Fig. 6*b*. It is evident that in the presence of the donor, CPM, the fluorescence intensity of the fluorescein at the 5′-end of the 20-mer is increased by ~220%.

Analogous experiments were performed with a 20-mer, $dT(pT)_{19}$-Fl-3′, having fluorescein located at the opposite 3′-end of the nucleic acid. Unlike the case of 5′-Fl-$dT(pT)_{19}$, formation of the complex with R14C-CPM causes only an ~8% decrease of the fluorescence of $dT(pT)_{19}$-Fl-3′ (λ_{ex} = 485 nm), which is the same as observed in the presence of unlabeled protein (data not shown). This difference results from the larger distance between CPM residues on the small 12-kDa domains of the DnaB hexamer and fluorescein at the 3′-end of the 20-mer (see below). The *dashed line* in Fig. 7*a* is the sum of the fluorescence emission spectra of independent components of the complex, R14C-CPM in the presence of unlabeled $dT(pT)_{19}$, and the fluorescence emission spectrum of $dT(pT)_{19}$-Fl-3′ in the presence of R14C-CPM (without energy transfer), in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP (λ_{ex} = 435 nm). The *solid line* in Fig. 7*a* is the fluorescence emission spectrum the R14C-CPM and $dT(pT)_{19}$-Fl-3′ complex at the same concentrations of the protein and nucleic acid as independent components of the complex. Contrary to the situation with 5′-Fl-$dT(pT)_{19}$, only a small difference is observed when both the donor, CPM on the DnaB protein, and the acceptor, fluorescein on the 3′-end of the 20-mer, are placed in the same complex as compared with the sum of the spectra of independent components of the complex. The emission intensity of R14C-CPM is only decreased by ~11% as compared with ~35% observed for R14C-CPM with 5′-Fl-$dT(pT)_{19}$, indicating a very diminished fluorescence energy transfer from CPM to the fluorescein moiety, when the acceptor is located at the 3′-end of the $dT(pT)_{19}$. Also, the sensitized emission of the fluorescein located at the 3′-end of the 20-mer is only increased by ~43% as compared with ~220% in the complex of R14C-CPM with 5′-Fl-$dT(pT)_{19}$ (Fig. 6*b*).

The dramatic difference between the emission spectrum of the complex of R14C-CPM with 5′-Fl-$dT(pT)_{19}$ and the spectrum of the complex with $dT(pT)_{19}$-Fl-3′ clearly shows that the helicase binds ssDNA in a predominantly single orientation, with respect to the polarity of the ssDNA sugar-phosphate backbone. If the helicase could bind ssDNA in two different orientations with equal probability, then the changes in the spectra of the complexes with the 20-mer, labeled with fluores-

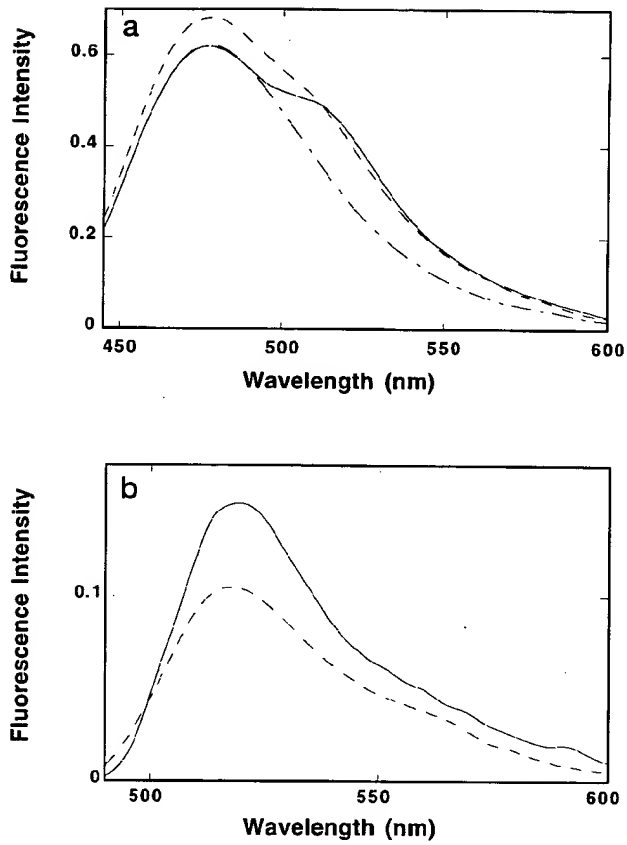


FIG. 7. *a*, sum of the fluorescence emission spectra (– – –) of R14C-CPM in the presence of unlabeled $dT(pT)_{19}$ (4.5×10^{-7} M (oligomer)) and $dT(pT)_{19}$-Fl-3′ in the presence of R14C-CPM (without energy transfer) (λ_{ex} = 435 nm) in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP and the fluorescence emission spectrum of the complex of R14C-CPM with $dT(pT)_{19}$-Fl-3′ (λ_{ex} = 435 nm) in the same solution conditions (——). Concentrations of $dT(pT)_{19}$-Fl-3′ and R14C-CPM are 4.5×10^{-7} M (oligomer) and 9.6×10^{-7} M (hexamer), respectively. The fluorescence emission spectrum of R14C-CPM normalized at 476 nm (peak) to the emission spectrum of R14C-CPM in the complex with $dT(pT)_{19}$-Fl-3′ (– · – · –) is also included. *b*, sensitized emission spectrum of $dT(pT)_{19}$-Fl-3′ (λ_{ex} = 435 nm) in the complex with R14C-CPM (——) obtained after subtraction of the normalized spectrum of R14C-CPM in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP superimposed on the fluorescence emission spectrum of $dT(pT)_{19}$-Fl-3′ in the presence of R14C-CPM (without energy transfer) (– – –), obtained at the same excitation wavelength. Concentrations of $dT(pT)_{19}$-Fl-3′ and R14C-CPM are 4.5×10^{-7} M (oligomer) and 9.6×10^{-7} M (hexamer), respectively.

cein at the 5′- or 3′-ends, would be indistinguishable.

The effect of the location of the fluorescence acceptor on the observed spectral properties of the studied complexes is reflected in the large differences in the true energy transfer efficiencies, E. Using Equations 5 and 8, we obtained the apparent transfer efficiencies of E_D = 0.77 ± 0.05 and E_A = 0.55 ± 0.03, respectively, for the complex of R14C-CPM with 5′-Fl-$dT(pT)_{19}$. This difference between E_D and E_A indicates that fluorescein, at the 5′-end of the bound $dT(pT)_{19}$, induces some additional nondipolar CPM fluorescence quenching. The true Förster fluorescence transfer efficiency from CPM, located on the small 12-kDa domain to the fluorescein residue at the 5′-end of $dT(pT)_{19}$, is then described by Equation 9, which yields E = 0.71 ± 0.05. Analogous calculations of the fluorescence energy transfer efficiency in the complex of R14C-CPM with $dT(pT)_{19}$-Fl-3′ yield E_D = 0.18 and E_A = 0.09 ± 0.01 (Table II). In this case, the true Förster transfer efficiency is E = 0.1 ± 0.01. The large difference between the true energy

TABLE II

Fluorescence properties of 5′-Fl-$dT(pT)_{19}$ and $dT(pT)_{19}$-Fl-3′ in the complex with the E. coli DnaB helicase and the DnaB helicase variant modified with CPM (R14C-CPM) in buffer T2 (pH 8.1, 10 °C) containing 100 mM NaCl and 1 mM AMP-PNP

Property	ssDNA oligomer	
	5′-Fl-$dT(pT)_{19}$	$dT(pT)_{19}$-Fl-3′
Fluorescence anisotropy (r)[a]	0.28 ± 0.01	0.24 ± 0.01
Limiting fluorescence anisotropy (r_o)[b]	0.29 ± 0.01	0.25 ± 0.01
Fluorescence energy transfer efficiency in the complex with R14C-CPM[c]	E_D = 0.77 ± 0.04 E_A = 0.55 ± 0.04 E = 0.71 ± 0.04	E_D = 0.18 ± 0.02 E_A = 0.09 ± 0.01 E = 0.1 ± 0.01

[a] λ_{ex} = 485 nm.
[b] λ_{ex} = 485 nm, determined using the Perrin equation (22).
[c] λ_{ex} = 435 nm.

transfer efficiencies shows that the 5′-end of the 20-mer, $dT(pT)_{19}$, is in much closer proximity to the CPM residues, which are located on the small domains of the DnaB hexamer, than to the 3′-end of the nucleic acid (see "Discussion").

The determination of exact distances between the donors (CPM) and acceptors (fluorescein) is beyond the scope of the present discussion on the mutual orientation between the DnaB helicase and the ssDNA in the complex. However, using Equation 11 we can estimate the approximate ratio of the distances between the 5′- and the 3′-end of the $dT(pT)_{19}$ oligomer from the center of the mass of CPM donors located on the small domains of the DnaB hexamer. Introducing E_1 = 0.71 and E_2 = 0.1 into Equation 11, we obtained R_1/R_2 = 0.60. Thus, the average distance of the 5′-end of the 20-mer is only 60% of the distance between the donors and the 3′-end of the nucleic acid.

Very similar behavior to the one described above has been observed when different donor-acceptor pairs have been used.[3] These results show, for the first time, that the DnaB hexamer binds ssDNA in a single orientation, with respect to the sugar-phosphate backbone of the nucleic acid. In the complex, the small 12-kDa and the large 33-kDa domains of the enzyme face the 5′- and 3′-ends of the nucleic acid, respectively.

DNA Mobility within the Strong and Weak DNA Binding Subsite of the DnaB Helicase—Assessment of the relative mobility of the different segments of the nucleic acid, within the DNA binding site, can be obtained by measuring the emission anisotropy of the fluorescent markers placed in different locations on the nucleic acid. To determine the relative mobility of ssDNA in two subsites of the total DNA binding site of the DnaB helicase, we determined the emission anisotropy of 5′-Fl-$dT(pT)_{19}$ and $dT(pT)_{19}$-Fl-3′ in the complex with the helicase. Anisotropies of both samples are constant across their emission spectra, indicating the lack of a significant local heterogeneity around the fluorescent markers (spectra not shown). However, the anisotropy of 5′-Fl-$dT(pT)_{19}$, r = 0.28 ± 0.01, is significantly higher than the anisotropy, r = 0.24 ± 0.01, determined for $dT(pT)_{19}$-Fl-3′. Because the fluorescence lifetimes of fluorescein in both complexes are very similar (~4 ns, data not shown), the obtained data indicate significantly higher mobility of the nucleic acid at its 3′-end.

Analogous fluorescence energy transfer and anisotropy studies with a 10-mer, $dA(pA)_9$, labeled with fluorescein at the 5′- or 3′-ends of the 10-mer indicate that its 5′-end is located in close proximity to the 12-kDa domain of the enzyme and has a similar strong decrease in its mobility (data not shown). As we described above, this oligomer binds exclusively to the strong subsite in the DNA binding site of the DnaB helicase. Thus,

[3] Jezewska, M. J., Rajendran, S., Bujalowska, and Bujalowski, W. (1998) *J. Biol. Chem.* **273**, in press.

fluorescence energy transfer and anisotropy data indicate that the nucleic acid binds with the first 10 nucleotides from its 5′-end to the strong DNA binding subsite of the total DNA binding site of the helicase.

DISCUSSION

The Total DNA Binding Site of a Helicase—Helicases play a key role in all aspects of DNA metabolism, and this role is related to the interactions of the enzyme with ssDNA and dsDNA controlled by binding and hydrolysis of a nucleoside triphosphate, *e.g.* ATP (30). Understanding the functional and structural aspects of the DNA binding site is a prerequisite for our understanding of how the enzymes perform their functions. Yet, little is known about the structure of the DNA binding site of any hexameric helicase and the functional interrelations within the binding site. In this work, we provide the first insight into the complex structure/function relationship of the DNA binding site of a hexameric replicative helicase, the *E. coli* DnaB protein.

Our previous studies with polymer ssDNA and ssDNA oligomers showed that in a stationary complex with the ATP-nonhydrolyzable analog, AMP-PNP, the enzyme has a single binding site located on a single subunit of the hexamer (12–14). Additionally, this single binding site is used when the enzyme binds to the DNA substrates resembling the replication fork (8, 13, 25, 26). These results indicate that the observed single binding site is, in fact, the total DNA binding site of the enzyme that, in functional complexes on the junction between ssDNA and dsDNA with the replication fork, encompasses both single- and double-stranded conformations of nucleic acid over a stretch of ~20 nucleotide residues.

The operational definition of the total binding site of the enzymes, which perform their catalysis on polymer lattices, such as helicases, should refer to the complex of the enzymes with a polymer substrate. A total binding site of an enzyme is used as a single entity that interacts with a continuous stretch of polymer substrate. This continuous fragment of the polymer substrate (DNA), within the total binding site, defines the site size of the enzyme-nucleic acid complex. The total binding site can be heterogeneous, *i.e.* built of functionally and/or structurally different areas, subsites, specific for the catalytic functions of the enzyme. However, the location of the subsites is sequential, *i.e.* they are placed along the polymer substrate. The total binding site can perform the dominant catalytic process characteristic for the enzyme, *e.g.*, unwinding of the duplex DNA. Such a binding site can be located on a single subunit of an oligomeric enzyme, such as the DnaB helicase; thus, there may be several total binding sites, but only one site (one subunit) at a time is engaged in interactions with DNA during the catalysis. A total binding site can include several subunits of an oligomeric enzyme, as in the case of DNA-dependent oligomeric polymerases.

Contrary to the total binding site of the enzyme, a subsite always interacts with a polymer DNA within the context of a total binding site. A subsite cannot be used as an independent entity in the interactions of the enzyme with polymer DNA; nor can it independently perform the catalysis.

The Total Binding Site of the DnaB Helicase Is Structurally Heterogeneous—Nuclease digestion protection studies provide a clear indication of the structural heterogeneity of the total binding site of the *E. coli* DnaB helicase. Only 10 or 11 nucleotide residues, within the total binding site, are strongly protected from digestion, while the remaining 9 or 10 residues are accessible to the nuclease (Fig. 1, *a* and *b*). These results indicate that the total binding site of the helicase, which occludes on ~20 nucleotide residues in the complex with polymer ssDNA, is built of two binding subsites each encompassing a similar number of ~10 nucleotides. Experiments on the binding of partial DNA ligands to the helicase showed a large difference in the affinities between the subsites and indicated that the 5′-end of the nucleic acid interacts with the strong binding subsite of the total binding site of the enzyme. The fact that the nuclease can access ~half of the total number of occluded residues within the entire binding site suggests not only a difference in the affinities between the subsites but also an open architecture of the hexamer at the subsite that encompasses the 3′-end of the nucleic acid (see below).

The Two DNA Binding Subsites of the Total Binding Site of the DnaB Helicase Have Dramatically Different Affinities for ssDNA—Direct evidence of large differences in the affinities between the DNA binding subsites of the DnaB helicase comes from the studies of the binding of a partial ligand, $d\epsilon A(p\epsilon A)_9$, to the enzyme. Using the thermodynamically rigorous method, we determined that only one 10-mer binds with significant affinity to the helicase and that the association is characterized by the binding constant $K_1 = (1.7 \pm 0.3) \times 10^7\ M^{-1}$. The affinity for the second binding subsite is characterized by $K_2 = (2.6 \pm 1) \times 10^4\ M^{-1}$; thus, it is ~3 orders of magnitude lower. It is evident that the major part of the free energy of binding of the helicase to ssDNA comes from interactions with the strong binding subsite. The very low affinity of the weak binding subsite indicates that the protein does not form efficient contacts with a single-stranded nucleic acid and suggests that this subsite of the helicase is not functionally a ssDNA binding site but rather that it fulfills a different role when the enzyme is in the complex with its physiological substrate, the replication fork (see below).

To efficiently form a complex with the strong DNA binding subsite, the nucleic acid must have a length of at least 6 or 7 nucleotide residues. No detectable affinities were observed with ssDNA oligomers shorter than 6 nucleotides in our solution conditions (Fig. 4*a*). It is interesting that the difference of 2 residues between 7- and 5-mer practically abolishes the affinity of the shorter oligomer for the binding site, while the same difference between the 8- and 10-mer leads to a decrease of the free energy of binding by only ~0.3 kcal/mol (Fig. 4*b*). A common misconception in studying protein-nucleic acid interactions is treating both a nucleic acid and a protein as interacting regular lattices. The difference between the free energy of interaction of oligomers of different lengths with the protein is then assigned to the difference in statistical effects between different oligomers, which usually has very poor quantitative justification. We point out that the nucleic acid is the only macromolecule that can be approximated by a regular lattice. The binding site on the protein can have a very complex structure, with distant regions making key contacts with the nucleic acid, hardly resembling a regular lattice. The differences between different oligomers in binding to the DnaB helicase cannot be explained by any difference in the statistical effect between the oligomers. This is particularly true for oligomers shorter than 6 or 7 residues. Rather, the results suggest that the elements of the strong binding subsite of the enzyme, which makes crucial binding contacts with the nucleic acid, are separated by a distance spanned by 6 or 7 nucleotides. The proper complex is formed only when all essential contacts are engaged in interactions with ssDNA. In this context, the similar number of ions released in the interactions of a 10-mer and a 6-mer with the helicase (~1.5) would be a result of the fact that a 6-mer can still form all essential contacts with the enzyme, although the oligomer constitutes only 60% of the length of the 10-mer (Fig. 5*b*).

Direct Proof That the DnaB Helicase Binds in a Single Orientation with Respect to the Sugar-Phosphate Backbone of a

ssDNA—Most of the studied helicases show preferential direction in the unwinding of dsDNA, *i.e.* in the 5′ → 3′ or the 3′ → 5′ direction (30). Therefore, it is often *a priori* assumed that the enzyme binds in strict polarity, 5′ → 3′ or 3′ → 5′, with respect to the orientation of the single-stranded nucleic acid strand. This is a natural assumption that simplifies current models, based on the still limited solution data, of how the helicase functions at the replication fork. However, one can argue that the enzyme can bind in both orientations to the nucleic acid lattice and that the proper orientation is imposed by specific interactions with dsDNA and/or multiple proteins that are building the machinery of the replication fork. In this context, it should be noted that several specific protein-protein interactions between the DnaB helicase and proteins, which are part of the primosome or replication fork complex, have been identified. Although recent electron microscopy and crystallographic data show polarity in a helicase binding to ssDNA (31, 32), the polarity in the binding of a helicase, with respect to the directionality of a ssDNA strand, has never been directly shown for any hexameric helicase in solution.

As we pointed out, the determination of the mutual orientation of the protein and nucleic acid in a complex should be based on the method that is sensitive to the differences in distances between different specific regions of both macromolecules. The fluorescence energy transfer technique is such a method. The difference in the effect of the location of the acceptor, fluorescein, at the 5′- or 3′-end of the 20-mer, $dT(pT)_{19}$, on the fluorescence spectra of the complex of nucleic acid with R14C-CPM (excited in a predominantly donor absorption band), is dramatic (Figs. 6 and 7). These dramatic spectral differences are reflected in the large differences between the energy transfer efficiencies from CPM in the small 12-kDa domains, all located at one end of the DnaB hexamer, and the fluorescein placed at the 5′- or 3′-end of the $dT(pT)_{19}$, which spans the entire DNA binding site. The efficiency, E, for the fluorescein placed at the 5′-end of the 20-mer is 0.71 ± 0.04. The efficiency of the same acceptor located at the 3′-end of the nucleic acid is only 0.10 ± 0.01. In the case of chemically identical donor-acceptor pairs, the transfer efficiency depends on two variable factors characteristic for the studied system, the distance between the donor and acceptor, R, and the orientation parameter, κ^2, which characterizes the mutual orientation of the donor absorption dipole and acceptor emission dipole (22). The value of κ^2 can theoretically assume any value between 0 and 4, but only these two extreme values would significantly affect the determined transfer efficiency. The possible range of κ^2 can be estimated by using the standard procedure (Ref. 24; Table II). The obtained ranges of κ^2 are very similar for both 5′-Fl-$dT(pT)_{19}$ and $dT(pT)_{19}$-Fl-3′ and away from the extreme values of 0 and 4 (Table II). Another equally rigorous procedure is to perform experiments with several different donor-acceptor pairs that, due to different structures of different chromophores, provide the necessary "randomization" of the orientations of emission and absorption dipoles (33). Using different donor-acceptor pairs, we obtained similar, very large differences between the fluorescence energy transfer efficiencies from the fluorophore on the 12-kDa domain of the DnaB helicase and the donor or acceptor placed at the 5′- and the 3′-end of the bound 20-mer (data not shown). The results clearly show that the large difference between the transfer efficiencies results from the large difference in the distances between the 5′-end and the 3′-end of the 20-mer and the CPM located on the small domain of the DnaB protomers.

Our data show that the DnaB helicase binds ssDNA in a predominately single orientation, with respect to the polarity of the single-stranded nucleic acid lattice. Moreover, the data show that, in the complex with ssDNA, the small domain of the protein is in close proximity to the 5′-end of the nucleic acid, while the large domain is located near the 3′-end of the bound ssDNA.

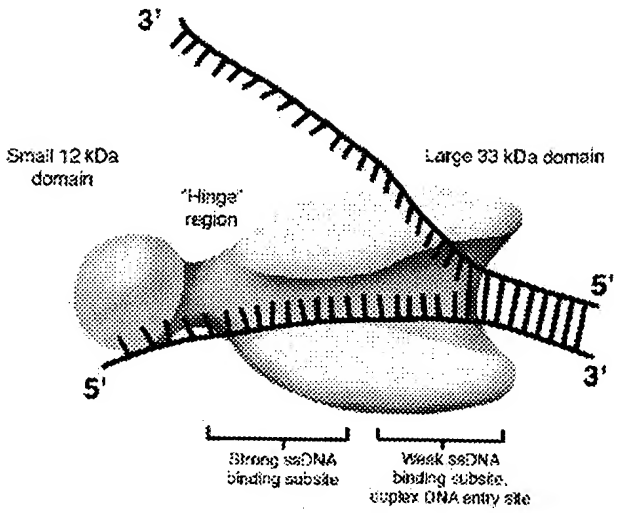


FIG. 8. **Schematic representation of the mutual orientation of the small 12-kDa and large 33-kDa domains of the single DnaB helicase protomer and the DNA binding subsites with respect to the polarity of the ssDNA, arms, and duplex DNA in the complex of the enzyme with replication fork, based on the results obtained in this work.** The helicase is preferentially bound to the 5′-arm of the replication fork using a single, total binding site of one of the six protomers shown in the figure. The locations of the DNA binding subsites within the total binding site is *sequential*. The weak binding subsite on the large 33-kDa domain faces the duplex part of the fork and constitutes the entry site for the dsDNA. The strong binding subsite is in the vicinity of the small 12-kDa domain and is engaged in interactions with the ssDNA at the 5′-end of the arm. The 3′-arm is not forming a stable complex with the helicase hexamer associated with the 5′-arm of the fork (26).

Sequential Locations of the Strong and Weak Binding Subsites of the DnaB Helicase—As determined in this work, the partial ligand, $d\epsilon A(p\epsilon A)_9$, binds with overwhelming preference to the strong binding subsite of the DnaB helicase. The transfer efficiency between the 5′-end of the 10-mer, $dA(pA)_9$, labeled with fluorescein, and CPM, located in the small domain of the DnaB protein, is very similar to the transfer efficiency between the CPM and fluorescein at the 5′-end of the 20-mer, $dT(pT)_{19}$ (data not shown). These results indicate that the 5′-ends of both the 10- and 20-mers are at a similar distance from the small domain of the protein. Thus, the strong ssDNA binding subsite encompasses the bound 20-mer at its 5′-end, which is in close proximity to the small 12-kDa domain of the protein, while the weak binding subsite is located entirely on the large domain. At present, it is unknown whether or not the small domain or the hinge region of the DnaB protomer are directly involved in interactions with ssDNA. It should be noted that the isolated large 33-kDa domain of the enzyme could still bind ssDNA with some affinity, although quantitative analysis of the binding has not been performed (28). Thus, it is possible that the small domain and the hinge region constitute a part of the strong ssDNA binding subsite of the intact DnaB helicase.

The DnaB helicase binds preferentially to the 5′-arm of the replication fork (26). Because the enzyme binds in a single orientation, with respect to the polarity of the ssDNA sugar-phosphate backbone, with the small domain facing the 5′-end of the ssDNA, it is evident that in the complex with the replication fork, the helicase hexamer is oriented with the large domains of the protomers toward the duplex part of the fork,

while the 5′-end of the arm of the replication fork is located in the vicinity of the small 12-kDa domains of the protomers. Anisotropy of the probe located at the 5′-end of the 20-mer bound to the helicase is significantly higher than the anisotropy of the same fluorescein residue located at the 3′-end of the nucleic acid (Table II). A significant decrease of the anisotropy, when the probe is located at the 3′-end of the bound nucleic acid, indicates an increased mobility of the nucleic acid in the weak binding subsite and is most probably due to the lack of strong contacts between the single-stranded nucleic acid and the binding site. Recall that the micrococcal nuclease can access the part of the nucleic acid in the weak subsite of the total binding site of the DnaB helicase, suggesting a more open structure of the total binding site at the 3′-end of the bound 20-mer.

The results described in this work provide an insight into the complex structure-function relationship within the DNA binding site of a replicative hexameric helicase. A model of the single, total DNA binding site on the DnaB protomer, engaged in the complex with the replication fork and based on the data presented in this work, is schematically shown in Fig. 8. The total DNA binding site of the enzyme is built of two subsites placed *sequentially* along the DNA substrate in the protein-nucleic acid complex. The strong ssDNA binding subsite occludes the 5′-end of the ssDNA, is located in close proximity to the small 12-kDa domain, and is distant from the duplex part of the fork. Binding of ssDNA to this subsite leads to the significant immobilization of the nucleic acid and provides the major part of the binding free energy. The subsite, which is located at the 3′-end of the ssDNA, binds the single-stranded nucleic acid very weakly. The single orientation of the helicase in the complex with ssDNA indicates that, when the enzyme approaches the replication fork, it faces the duplex part of the fork with the weak binding subsite located entirely on the large 33-kDa domain of the protein. Thus, the weak binding subsite constitutes the entry site for the dsDNA in the fork. The more open architecture of this subsite provides a large space, which is necessary for the incoming duplex DNA.

Comparison with other hexameric helicases is difficult because, at this time, no analogous data on the structure of their nucleic acid binding sites are available. However, it is possible that similar functional and structural relationships within the DNA binding site are general for all other hexameric helicases.

Acknowledgments—We thank Dr. T. Wood from the NIEHS Center, National Institutes of Health, for excellent work in obtaining the DnaB protein variant R14C and for many helpful discussions. We thank Gloria Drennan Davis for help in preparing the manuscript.

REFERENCES

1. Kornberg, A., and Baker, T. A. (1992) *DNA Replication*, W. H. Freeman and Co., San Francisco
2. Wickner, S., Wright, M., and Hurwitz, J. (1973) *Proc. Natl. Acad. Sci. U. S. A.* **71,** 783–787
3. McMacken, R., and Kornberg, A. (1977) *J. Biol. Chem.* **253,** 3313–3319
4. Ueda, K., McMacken, R., and Kornberg, A. (1978) *J. Biol. Chem.* **253,** 261–269
5. LeBowitz, J. H., and McMacken, R. (1986) *J. Biol. Chem.* **261,** 4738–4748
6. Baker, T. A., Funnell, B. E., and Kornberg, A. (1987) *J. Biol. Chem.* **262,** 6877–6885
7. Bujalowski, W., Klonowska, M. M., and Jezewska, M. J. (1994) *J. Biol. Chem.* **269,** 31350–31358
8. Jezewska, M. J., and Bujalowski, W. (1996) *J. Biol. Chem.* **271,** 4261–4265
9. Reha-Krantz, L. J., and Hurwitz, J. (1978) *J. Biol. Chem.* **253,** 4051–4057
10. San Martin, M. C., Valpuesta, J. M., Stamford, N. P. J., Dixon, N. E., and Carazo, J. M. (1995) *J. Struct. Biol.* **114,** 167–176
11. Yu, X., Jezewska, M. J., Bujalowski, W., and Egelman, E. H. (1996) *J. Mol. Biol.* **259,** 7–14
12. Bujalowski, W., and Jezewska, M. J. (1995) *Biochemistry* **34,** 8513–8519
13. Jezewska, M. J., Kim, U-S., and Bujalowski, W. (1996) *Biochemistry* **35,** 2129–2145
14. Jezewska, M. J., Kim, U-S., and Bujalowski, W. (1996) *Biophys. J.* **71,** 2075–2086
15. Bujalowski, W., and Klonowska, M. M. (1993) *Biochemistry* **32,** 5888–5900
16. Bujalowski, W., and Klonowska, M. M. (1994) *Biochemistry* **33,** 4682–4694
17. Bujalowski, W., and Klonowska, M. M. (1994) *J. Biol. Chem.* **269,** 31359–31371
18. Secrist, J. A., Barlo, J. R., Leonard, N. J., and Weber, G. (1972) *Biochemistry* **11,** 3499–3506
19. Ledneva, R. K., Razjivin, A. P., Kost, A. A., and Bogdanov, A. A. (1977) *Nucleic Acid Res.* **5,** 4226–4243
20. Jezewska, M. J., and Bujalowski, W. (1997) *Biophys. Chem.* **64,** 253–269
21. Azumi T., and McGlynn, S. P. (1962) *J. Chem. Phys.* **37,** 2413–2420
22. Lakowicz, J. R. (1983) *Principle of Fluorescence Spectroscopy*, pp. 111–339, Plenum Publishing Corp., New York
23. Berman, H. A, Yguerabide, J., and Taylor, P. (1980) *Biochemistry* **19,** 2226–2235
24. Dale, R. E., Esinger, J., and Blumberg, W. E. (1979) *Biophys. J.* **26,** 161–194
25. Jezewska, M. J., and Bujalowski, W. (1996) *Biochemistry* **35,** 2117–2128
26. Jezewska, M. J., Rajendran, S., and Bujalowski, W. (1997) *Biochemistry* **36,** 10320–10326
27. Record, M. T., Lohman, T. M., and deHaseth, P. L. (1976) *J. Mol. Biol.* **107,** 145–158
28. Nakayama, N., Arai, N., Bond, M. N., Kaziro, Y., and Arai, K. (1984) *J. Biol. Chem.* **259,** 97–101
29. Heyduk, T., and Lee, J. C. (1994) *Biochemistry* **31,** 5165–5171
30. Lohman, T. M., and Bjornson, K. P. (1996) *Annu. Rev. Biochem.* **65,** 169–214
31. Egelman, E. H., Yu, X., Wild, R., Hingorani, M. M., and Patel, S. S. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92,** 3869–3873
32. Korolev, S., Hsieh, J., Gauss, G. H., Lohman, T. M., and Waksman, G. (1997) *Cell* **90,** 635–647
33. Cantor, C. R., and Pechukas, P. (1971) *Proc. Natl. Acad. Sci. U. S. A.* **68,** 2099–2101

EXHIBIT 5

NCBI **BLAST 2 sequences** BLAST Entrez ?

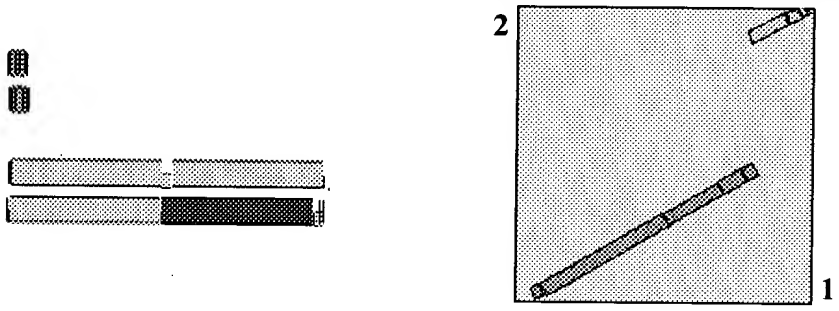
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.3 [Apr-24-2002]

Matrix BLOSUM62 gap open: 11 gap extension: 1

x_dropoff: 50 expect: 10.00 wordsize: 3 Filter Align

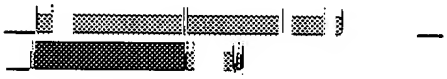
Sequence 1 lcl|9306aa.txt **Length** 546 (1 .. 546)

Sequence 2 gi 2661771 helicase [Rhodothermus marinus] **Length** 945 (1 .. 945)



NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  350 bits (899), Expect = 1e-95
Identities = 186/407 (45%), Positives = 270/407 (65%), Gaps = 8/407
(1%)
```



```
Query: 36   RRNTRSTAKSKVQPVNDYGRIQPQAPELEEAVLGALMIEKDAYSLVSEILRPESFYEHRH
95

             RR TR+   +  Q     GR+ PQA ELE+AVLGA++IE +A     EIL PE+FY+ RH
Sbjct: 29   RRRTRAQIHALHQQA---GRVPPQAVELEQAVLGAMLIEPEAIPRALEILTPEAFYDGRH
85

Query: 96   QLIYAAITDLAVNQKPVDILTVKEQLSKRGELEEVGGPFYITQLSSKVASSAHIEYHARI
155

             Q I+ AI  L    + VD+LTV E+L + GELE+ G    Y+++L+++VAS+A++EYHARI
Sbjct: 86   QRIFRAIVRLFEQNRGVDLLTVTEELRRTGELEQAGDTIYLSELTTRVASAANVEYHARI
145
```

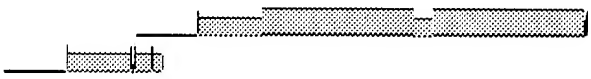
```
Query: 156 IAQKYLARELITFTSNIQSKAFDETLDVDDLMQEAEGKLFEISQRNMKKDYTQINPIIAE
215
           IA+K L R +I   + +  +A+D   D  +L+ E E ++F +S  +++K    +N ++ E
Sbjct: 146 IAEKALLRRMIEVMTLLVGRAYDPAADAFELLDEVEAEIFRLSDVHLRKAARSMNEVVKE
205

Query: 216 AYEQIQKAAARTDGLSGLESGYTKLDKMTSGWQKSDLIIIAARPAMGKTAFVLSMAKNIA
275
             E+++     R  G++G+ SG+ +LD +T GWQ+ DLIIIAARP+MGKTAF LS A+N A
Sbjct: 206 TLERLEAIHGRPGGITGVPSGFHQLDALTGGWQRGDLIIIAARPSMGKTAFALSCARNAA
265

Query: 276 V--NFRNPVALFSLEMSNVQLVNRLISNVCEIPSEKIKSGQLAAYEWQQXXXXXXXXXXX
333
           +  ++   VA+FSLEM   QL  RL++    + ++  ++G+L   +W++
Sbjct: 266 LHPHYGTGVAIFSLEMGAEQLAQRLLTAEARVDAQAARTGRLRDEDWRKLARAAGRLSDA
325

Query: 334 XXXVDDTPSLSVFELRTKARRLVREHGVRIIIIDYLQLMNASGM-AFGSRQEEVSTISRS
392
              +DDTPSL V ELR K RRL  EH + ++I+DYLQLM AS M    +R++E++ ISRS
Sbjct: 326 PIFIDDTPSLGVLELRAKCRRLKAEHDIGLVIVDYLQLMQASHMPRNANREQEIAQISRS
385

Query: 393 LKGLAKELNIPIIALSQLNRGVESREGLEGKRPQLSDLRESGAIEQD 439
           LK LAKELN+P++ALSQL+R VE+R G   KRPQLSDLRESG +  D
Sbjct: 386 LKALAKELNVPVVALSQLSRAVETRGG--DKRPQLSDLRESGCLAGD 430
Score = 76.6 bits (187), Expect = 5e-13
Identities = 44/103 (42%), Positives = 59/103 (56%), Gaps = 5/103 (4%)
```



```
Query: 428 SDLRESGAIEQDADMVCFIHRPEYYKIFQDDKGNDLRGMAEIIIAKHRNGAVGDVLLRFK
487
           +D+    +IEQDAD+V FI+RPE Y I  D+ GN   G+AEIII K RNG  G V L F
Sbjct: 847 NDIIAHNSIEQDADVVLFIYRPERYGITVDENGNPTEGIAEIIIGKQRNGPTGTVRLAFI
906

Query: 488 GEYTRFQNPDDDMVIPLPDAGAMLGSRMNNTGTVPPPPAEFAP 530
            +Y RF+N    + +  P+ G  L    + T  +P  P + AP
Sbjct: 907 NQYARFEN----LTMYQPEPGTPLPETPDET-ILPSGPPDEAP 944
CPU time:     0.08 user secs.     0.03 sys. secs         0.12 total
secs.

Lambda      K       H
   0.317    0.134    0.378

Gapped
Lambda      K       H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
```

Number of Hits to DB: 4104
Number of Sequences: 0
Number of extensions: 321
Number of successful extensions: 5
Number of sequences better than 10.0: 1
Number of HSP's better than 10.0 without gapping: 1
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 0
Number of HSP's gapped (non-prelim): 2
length of query: 546
length of database: 181,542,687
effective HSP length: 124
effective length of query: 422
effective length of database: 140,313,307
effective search space: 59212215554
effective search space used: 59212215554
T: 9
A: 40
X1: 16 (7.3 bits)
X2: 129 (49.7 bits)
X3: 129 (49.7 bits)
S1: 41 (21.7 bits)
S2: 73 (32.7 bits)

Protein Engineering vol.12 no.2 pp.85–94, 1999

Twilight zone of protein sequence alignments

Burkhard Rost[1,2,3]

[1]EMBL, 69 012 Heidelberg, [2]LION Bioscience AG, Im Neuenheimer Feld 517, 69 120 Heidelberg, Germany and [3]Columbia University, Department of Biochemistry and Molecular Biophysics, 650 West 168 Street, New York, NY 10032, USA

Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20–35% sequence identity. Here, more than a million sequence alignments were analysed between protein pairs of known structures to re-define a line distinguishing between true and false positives for low levels of similarity. Four results stood out. (i) The transition from the safe zone of sequence alignment into the twilight zone is described by an explosion of false negatives. More than 95% of all pairs detected in the twilight zone had different structures. More precisely, above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were. (ii) Whether or not sequence homology implied structural identity depended crucially on the alignment length. For example, if 10 residues were similar in an alignment of length 16 (>60%), structural similarity could not be inferred. (iii) The 'more similar than identical' rule (discarding all pairs for which percentage similarity was lower than percentage identity) reduced false positives significantly. (iv) Using intermediate sequences for finding links between more distant families was almost as successful: pairs were predicted to be homologous when the respective sequence families had proteins in common. All findings are applicable to automatic database searches.

Keywords: alignment quality analysis/evolutionary conservation/genome analysis/protein sequence alignment/sequence space hopping

Introduction

Protein sequence alignments in twilight zone

Protein sequences fold into unique three-dimensional (3D) structures. However, proteins with similar sequences adopt similar structures (Zuckerkandl and Pauling, 1965; Doolittle, 1981; Doolittle, 1986; Chothia and Lesk, 1986). Indeed, most protein pairs with more than 30 out of 100 identical residues were found to be structurally similar (Sander and Schneider, 1991). This high robustness of structures with respect to residue exchanges explains partly the robustness of organisms with respect to gene-replication errors, and it allows for the variety in evolution (Zuckerkandl and Pauling, 1965; Zuckerkandl, 1976; Doolittle, 1979, 1986). Structure alignments have uncovered homologous protein pairs with less than 10% pairwise sequence identity (Valencia *et al.*, 1991; Holmes *et al.*, 1993; Holm and Sander, 1996; Brenner *et al.*, 1996; Hubbard *et al.*, 1997). Indeed, most similar protein structure pairs appear to have less than 12% pairwise sequence identity (Rost, 1997). Furthermore, the average sequence identity between all pairs of similar structures is supposedly 8–10%, and the observed distribution (Gaussian peaking around 8% identity) marks another region, the midnight zone (Rost, 1997). The midnight zone is populated by protein structure pairs that may have become similar by convergent or divergent evolution (Doolittle, 1994; Rost, 1997). Threading algorithms ultimately aim at revealing homologous pairs from the midnight zone (Wodak and Rooman, 1993; Bryant and Altschul, 1995; Sippl, 1995; Rost and Sander, 1996; Sippl and Floeckner, 1996; Fischer *et al.*, 1996; Rost and O'Donoghue, 1997). Conventional sequence alignment methods become problematic at much higher values of sequence identity. Methods often fail to correctly align protein pairs with 20–30% pairwise sequence identity. Hence, Doolittle (1986) coined the term twilight zone for sequence alignments in this region. Do the difficulties of alignment methods in this zone reflect merely technical difficulties (statistical significance of detection), or is the twilight zone defined by a particular feature of evolution?

Length-dependent cut-off for significant sequence identity

Pairwise sequence identity (percentage of residues identical between two proteins) is not sufficient to define the twilight zone. Instead, analysing the relatively small number of structure pairs available in 1990, Sander and Schneider (1991) defined a length-dependent threshold for significant sequence identity. The threshold curve defined (dubbed HSSP-curve) was roughly proportional to the inverse square-root of the length for alignments between 7 and 80 residues, and was clipped to saturate at 25% sequence identity over more than 80 residues. In 1990, no pair with more than 30 identical residues of 100 aligned had different structures (Sander and Schneider, 1991). Was this still true for the five times larger PDB (Bernstein *et al.*, 1977) of 1997?

Hopping in sequence space

If we could plot the space of protein sequences, would we observe the protein families as islands? Unfortunately, we cannot tell. Nevertheless, useful information has been extracted from sequence (Casari *et al.*, 1995) and structure (Maiorov and Crippen, 1995) space. In everyday database searches, protein families are widened by exploiting the transitivity of homology (Pearson, 1996): (i) a query sequence U is aligned to a database, say SWISS-PROT (Bairoch and Apweiler, 1997); (ii) all sequences aligned at levels of significant similarity are used as new seeds U_i, and for each U_i SWISS-PROT is searched again; (iii) this procedure is repeated until no new sequences are found. Sequence space hopping may be used in combination with knowledge from structures to widen families (Holm and Sander, 1997), or to increase the information contained in multiple sequence alignments input to prediction methods (Rost, 1996, 1997). Recently, the transitivity of protein families has been exploited successfully to automatically increase the yield in database searches [Ruben Abagyan

presented the 'multi-link recognition' method 1996 at the CASP2 meeting (Abagyan and Batalov, 1997); Park *et al.* (1997) presented the 'intermediate sequence search' method and Neuwald *et al.* (1997) implemented the same concept (Neuwald, *et al.*, 1997)]. Here, I confirm the original findings based on a different data set, and analysed in detail how the gain depended on the number of intermediate sequence, and their similarity.

Here, I present results of aligning a set of 792 sequence-unique (no pair in set has more than 25% sequence identity) proteins of known structure against PDB. The following questions were investigated. Is the number of protein pairs of non-similar structures proportional to the distance from the HSSP-curve (eqn 1), or do false positives increase more rapidly in the twilight zone? Is the curve defined by Sander and Schneider (1991) still valid? Would using sequence similarity rather than identity improve accuracy (as speculated by Schneider and Sander)? Finally, can the accuracy be improved for pair alignments by expert rules? The results verify, partially, earlier work based on a 1000-fold larger data set (Sander and Schneider, 1991). The novel aspects were (i) a definition of a threshold for similarity (eqn 2), and a refinement of the threshold for identity; (ii) an introduction of various expert rules. Aspects largely complementing other analyses were (Abagyan and Batalov, 1997; Park *et al.*, 1997; Brenner *et al.*, 1998): (i) a large-scale evaluation of exploiting intermediate sequences (sequence-space-hopping); (ii) a detailed analysis of true and false positives providing estimates for accuracy and coverage of database searches; and (iii) a comparison with BLAST, one of the most popular methods for rapid databases searches (Altschul *et al.*, 1990; Altschul and Gish, 1996).

Methods

Data set: 792 sequence-unique protein structures

Protein databases are biased towards particular protein families. To reduce this bias, analyses are usually restricted to representative data sets (Hobohm *et al.*, 1992). Here, I chose the maximal set of sequence-unique proteins of known structure available in early 1997 (Holm and Sander, 1996). 'Sequence-unique' was defined as 'no pair in the set falls above the HSSP-curve (eqn 1; Sander and Schneider, 1991). As a rule-of-thumb, no pair had more than 25% pairwise sequence identity. Each of these proteins was aligned against the subset of PDB contained in the early 1997 release of the FSSP database of protein structure alignments (Holm and Sander, 1996). This subset amounted in total to about 5646 protein chains. Obviously the second step (792 versus 5646) re-introduced bias into the results. However, aligning the 792 sequence-unique pairs against themselves would not have yielded any result for most of the twilight zone analysed here. Thus, 792 versus 5646 was the best compromise in reducing bias *and* monitoring the biased region. The resulting test set was the largest possible set of proteins for which structural information was available (and thus false and correct hits could be automatically distinguished).

Generation of sequence alignments

Protein pairs were aligned by two different program types. (i) Full dynamic programming as implemented in the Smith–Waterman (Smith and Waterman, 1981) based method MaxHom (Schneider, 1994) (McLachlan metric, with minimum = –0.5, maximum = 1.00, and gap open = 3, gap elongation = 0.3); and (ii) quick database searches as implemented by the two versions of the BLAST series: BLASTP (Altschul *et al.*, 1990; Altschul and Gish, 1996), and PSI-BLAST (Altschul *et al.*, 1997). All 792 unique proteins were aligned against all 5646 proteins from the PDB subset. Alignments shorter than 10 residues were not considered, as identical polypeptides of up 10 residues are known to occur in different structure states (Kabsch and Sander, 1984; Cohen *et al.*, 1993). Technical limitations (CPU time) required the restriction of the dynamic-programming analysis to the best 2000 hits for each of the 792 unique proteins. (Note: this restriction applied only to the final displayed alignment. Of course, all possible combinations were explored initially by the alignment algorithm.) The resulting final data set comprised about 1.7 million pairwise alignments. For the comparison between the dynamic programming and the BLAST methods, the data set had to be reduced to all pairs that were aligned by all methods compared (the problem was that neither BLASTP, nor PSI-BLAST could be forced to report absolutely wrong, i.e. ALL pairwise alignments).

Definition of sequence identity and sequence similarity

(i) Pairwise sequence identity was defined by the percentage of residues identical between two aligned sequences (e.g. aspartic matching aspartic counts 1: D – D = 1; aspartic on glutamic was a non-match: D – E = 0). (ii) Pairwise sequence similarity was defined by the percentage of residues similar between two sequences (e.g. D – D $\leq$ 1; and aspartic on glutamic was now considered a match: D – E > 0). Similarity scores depend on the particular metric used to capture physico-chemical properties of amino acids (note: most amino acids are not considered 100% similar to themselves by typical metrices, as such metrices are based on log-odds, e.g. for the McLachlan metric only F, W, Y and C yield 100% self-similarity). Consequently, levels of similarity are not directly comparable between different metrices. For comparability, I used the McLachlan metric (Gribskov *et al.*, 1987) also used in the HSSP database (Schneider *et al.*, 1997). In principle, there are two ways to convert similarity into percentage values: (i) by normalizing the similarity score by the maximal possible score observed in a given metric (percentage residue similarity); and (ii) by setting an arbitrary threshold of the similarity score to distinguish similar–not similar and counting the percentage of residues that are similar according to this threshold (percentage of similar residues). Again, I followed the practice of the HSSP database compiling the percentage residue similarity (normalized by maximal possible scores). When compiling percentages, the number of identical residues was normalized by the number of residues aligned, gaps were ignored.

Standard of truth for structural similarity

Similarity between two protein structures is not uniquely defined. Different structure alignment methods yield different scores (Alexandrov *et al.*, 1992; Holm *et al.*, 1993; Luo *et al.*, 1993; Orengo, 1994; Crippen and Maiorov, 1995; Gerstein and Levitt, 1996; Holm and Sander, 1996; Orengo and Taylor, 1996; Zu-Kang and Sippl, 1996). Such differences can be substantial, as illustrated by differences between the expert-based database of structural alignments SCOP (Murzin *et al.*, 1995; Brenner *et al.*, 1996; Hubbard *et al.*, 1997), and the automatically generated databases CATH (Orengo *et al.*, 1993, 1997) and FSSP (Holm and Sander, 1996). In general, FSSP tends to find more pairs of similar structure than do CATH and SCOP. However, this is only a trend. For many examples, SCOP finds structural similarity and FSSP does not. Here, I

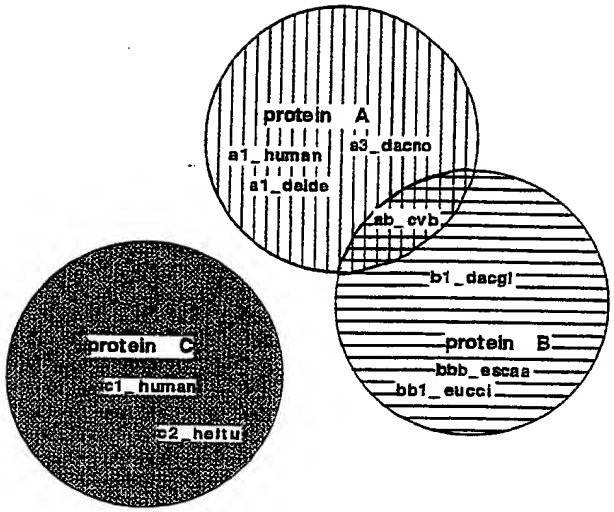


Fig. 1. Sketch of sequence-space-hopping. The triangle defines three search proteins (A, B and C) having mutually less than 25% sequence identity. The circles define the three families (all sequences inside the circle indicated by arbitrary names *aaa_species* have more than 25% sequence identity to the respective search proteins A, B and C). Sequence-space-hopping implies joining the circles representing the protein families (as shown for proteins A and B in the striped circles) if they contain identical proteins that are aligned in the same region (*ab_cvb* in the example given).

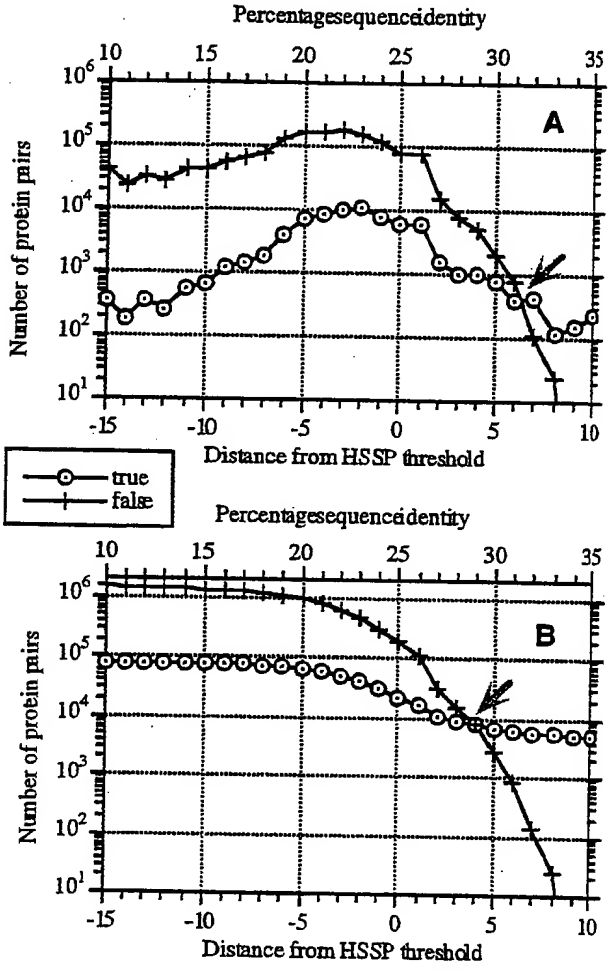


Fig. 2. Explosion of structurally dissimilar pairs in the twilight zone. Numbers of true (pairs with similar structure) and of false positives (pairs with no similar structure) plotted versus the distance to the HSSP-curve (Sander and Schneider, 1991), i.e. the horizontal axes give the distance from the threshold defined in eqn 1 (numbers refer to the parameter n in eqn 1). The levels of pairwise sequence identity corresponding to the distance were shown on top. (**A**) Number of pairs observed at any distance (logarithmic scale). (**B**) Cumulative number of pairs observed (logarithmic scale). For example, at a threshold corresponding to about 32% sequence identity for long alignments, the numbers of true and false positives were equal (arrow in A); at about 29% even the cumulative numbers of true and false positives were equal (arrow in B). Note: numbers of true negatives and false negatives result from the cumulative sums left of the threshold; percentages of true and false positives given in Figure 5.

chose the FSSP database 'a standard of truth': any pair for which FSSP listed a significant score [zDALI > 4 (Holm and Sander, 1996)] of structural similarity was considered to be structurally similar. In order to distinguish between true and false positives this decision implied that all pairs not listed at the given cut-off of the FSSP database were structurally not similar. However, this brought up the problem of different structure alignment methods. For example SCOP may consider a pair structurally similar, and FSSP may not. Thus, additionally all pairs were excluded from the analysis that were listed in FSSP but with lower z-scores. Even that still left pairs of proteins with clear levels of sequence identity (more than 40%) which were not found listed in FSSP. Thus, I had to refine this procedure by semi-automatically checking the structural similarity for about 2000 protein pairs all of which had levels of above 30% pairwise sequence identity [note this number was negligibly small, as only 1% of all pairs were found above this value (Fig. 2B)!]. The particular way in which the standard-of-truth was constructed implied that estimates for true positives might be slightly optimistic, estimates for false negatives slightly pessimistic.

Concept of true and false hits

When Chothia and Lesk (1986) first analysed the relation between sequence and structure similarity, they monitored the details of structural differences, and found that the differences are inversely proportional to the level of sequence identity. The binary notion of 'similar structure' (true or false) used in this analysis reflected a different focus: the goal was to estimate the accuracy in correctly detecting rather than in correctly aligning homologues. Did this imply that correct detection and correct alignment were not correlated (as often the case for threading: Bryant and Altschul, 1995; Lemer *et al.*, 1995; Sippl, 1995; Fischer *et al.*, 1996)? Not necessarily, but the fact is that two homologues can be detected although part—or even the entire—alignment is wrong. (However, this extremely irritating point was not pursued further in this analysis.) The following cases were distinguished: (i) true positives, alignments between proteins of similar structure that fall above a given threshold (defined by the sequence alignment method); (ii) false positives, alignments between proteins of dissimilar structure that fall above a given threshold of the sequence alignment; (iii) true negatives, alignments between proteins of dissimilar structure that fall below a given threshold; and (iv) false negatives, alignments between proteins of similar structure that fall below a given threshold. Note that 'negatives' and 'positives' represent two sides of the same coin: at

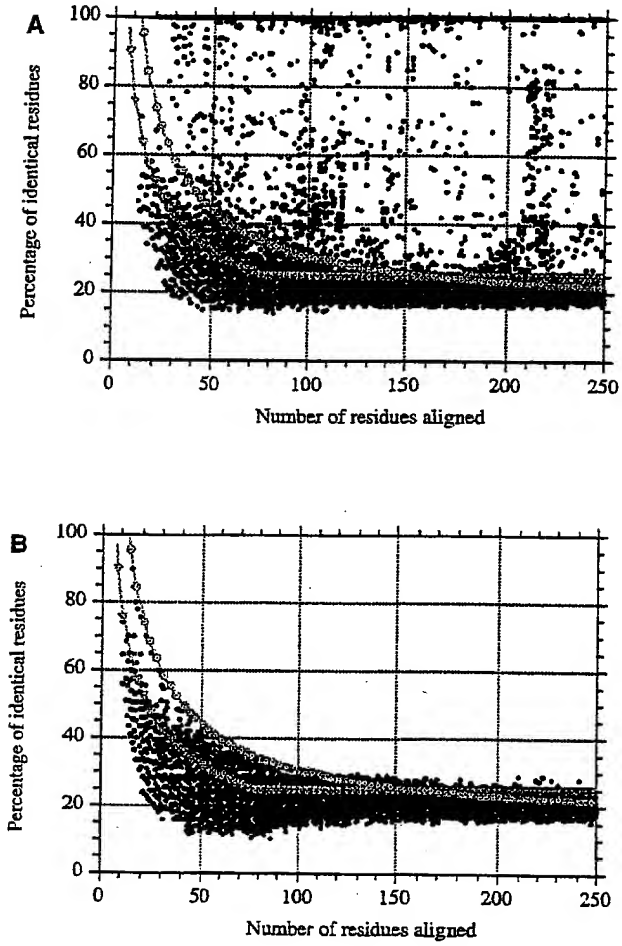


Fig. 3. Pairwise sequence identity versus alignment length. The original HSSP-curve (Sander and Schneider, 1991) (dotted circles, eqn 1) appeared to fit the true positives (homologues, A) better than the false positives (B). In contrast, the new curve proposed here (filled diamonds, eqn 2) was more conservative in excluding false positives. Note that due to the huge number of pairs the plots for true (A) and false (B) positives appeared almost equally densely populated (Figure 2 revealed the problem of such a scatter plot).

any threshold extracted from the sequence alignment n, the following equations hold (for cumulative numbers):

false negatives + true positives = all pairs of similar structure

true negatives + false positives = all pairs of dissimilar structure.

Distance to HSSP threshold

The HSSP-curve was originally defined by (Sander and Schneider, 1991):

$$p^I(n) = n + \begin{cases} 290.15 \cdot L^{-0.562}, & \text{for } L < 80 \\ 25, & \text{doe } L \geq 80 \end{cases} \quad (1)$$

where L gave the number of residues aligned between two proteins; p^I the cut-off percentage of identical residues over the L aligned residues; and n described the distance in percentage points from the curve ($n = 0$ corresponds to the original HSSP-curve; $n = 5$ to the official HSSP database releases; curve plotted in Figure 3). Once Schneider and Sander (1991) had discovered the basic functional dependence between sequence identity and alignment length, they merely had to fix two free parameters: the factor and the exponent. Both were chosen to fit the data observed in 1991, in particular to reach values of 25% around alignment length of 80, and values of 100% around alignment length of 10. The principle functional dependence described by eqn 1 also follows from statistics, as was recently shown in an elegant work (Alexandrov and Soloveyev, 1998). Let p_i (i = 1,..., 20) be the probability that amino acid i occurs in a protein, and m_{ij} the score for randomly aligning two amino acids i and j. The score S of an entire alignment can then be approximated by:

$$S = \langle m \rangle \cdot L$$

where $\langle m \rangle$ is the expectation value of m_{ij}, and L the alignment length. If the values of m_{ij} are independent, Gaussian distributed variables, it follows (after some elementary operations) that the relation between the standard deviation of the values of m_{ij} (σ_m), and the resulting score distribution (σ_S) is:

$$\sigma_m = L^{-0.5} \cdot \sigma_s$$

In their original article Alexandrov and Soloveyev work out an appropriate re-scaling of the dynamic programming alignment. However, this scheme cannot be applied after the alignment has been completed (as the threshold functions used in this work), rather it has to be implemented into the alignment method.

New curve for length-dependent significance of pairwise sequence identity

I attempted to solve the problems of the original HSSP-curve (eqn 1; Results) by defining the following curve for the separation of true and false positives (Figure 3, grey line with dotted circles):

$$p^I(n = n + 480 \cdot L^{-0.32 \cdot (1 + e^{-L/1000})} \quad (2)$$

where L gave the number of residues aligned between two proteins; p^I the cut-off percentage of identical residues over the L aligned residues; and n described the distance in percentage points from the curve ($n = 0$ plotted in Figure 3). The constraints in visually selecting the final function were (i) to maintain the functional form defined by eqn 1 (and suggested by the statistics of Alexandrov and Soloveyev, 1998); (ii) to hit the 100% mark at alignments that are too short to reveal anything about structural similarity (= 11 residues); (iii) to saturate at levels around 20% sequence identity (reached for length = 300); and (iv) to roughly reflect the observed gradient. Saturation for long alignments was realized by the functional form of the exponent (note: the term + $e^{-L/a}$ resulted in an exponential decay). This 'saturation' constraint also afflicted the particular value of the factor (0.32 rather than about 0.5 as suggested by the distribution of the data, Figure 4).

New curve for length-dependent significance of pairwise sequence similarity

The original HSSP-curve was derived for sequence identity, not for sequence similarity (Sander and Schneider, 1991). The functional dependence between similarity and length appeared comparable to the one between identity and length (Results). This prompted a similar definition for the separation between true and false positives based on similarity:

$$p^S(n = n + 420 \cdot L^{-0.335 \cdot (1 + e^{-L/2000})} \quad (3)$$

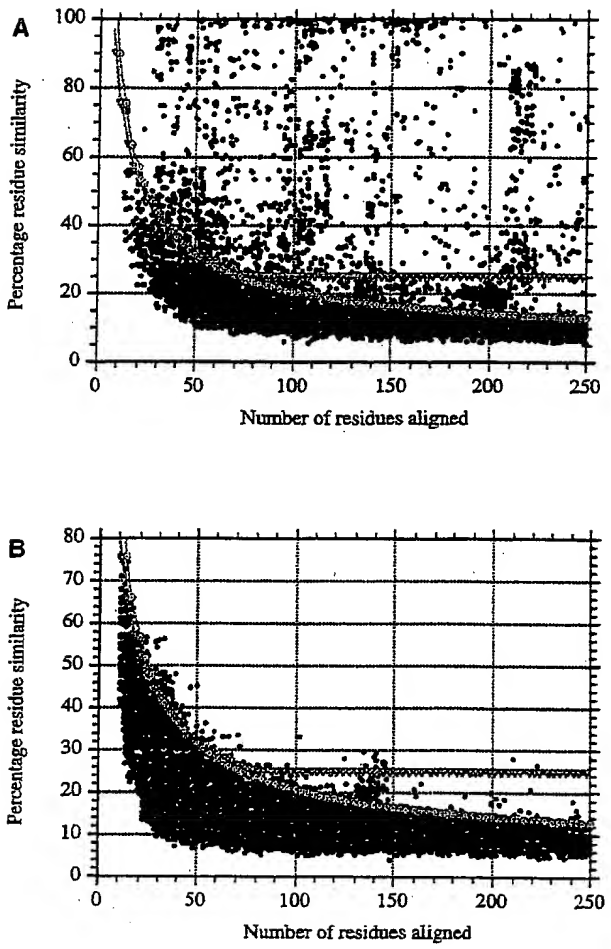


Fig. 4. Pairwise sequence similarity versus alignment length. (**A**) Correctly detected structural homologues; (**B**) false positives. Open circles, original HSSP-curve (Sander and Schneider, 1991) (eqn 1); filled triangles, new curve proposed here (eqn 3).

where L gave the number of residues aligned between two proteins; p^S defined cut-off for the percentage of residue similarity over the L aligned residues; and n described the distance in percentage points from the curve (n = 0 plotted in Figure 4).

Sequence-space-hopping

Suppose proteins A_0 and B_0 were less than 25% identical; family A is given by: $\{A_0, A_1,..., A_n\}$ (such that all proteins in the family A are more than 25% identical to A_0); analogously family B is given by: $\{B_0, B_1,..., B_m\}$. Although A_0 and B_0 differed by more than 75%, it may well be true that both were aligned to the same sequences, i.e. that for some i and j: $A_i = B_j$. If this is the case, 'sequence-space-hopping' refers to simply extending both families A and B to become: $\{A_0, A_1,..., A_n, B_0, B_1,..., B_m\}$ (Figure 1). Technically, I described this situation by compiling a simple matrix $H(A,B)$ that contained the number of overlapping proteins (i.e. those contained both in family A and B) between all proteins in the test set (792 chains) and all proteins in the search set (5646 chains). For example, $H(A,B) = 5$ implied that test protein A and search protein B had five identical proteins in their family alignments. The family alignments were taken from the HSSP database (Schneider *et al.*, 1997) with a cut-off at: HSSP-curve + 10% (n = 10 in eqn 1), i.e. for alignments longer than 80 residues, 35% pairwise sequence identity was required. All protein pairs (A,B) in the twilight zone were investigated for which $H(A,B)$ was larger than zero. Note, the concept of sequence-space-hopping explored here is being used in everyday sequence analysis. The novel idea introduced by others (Abagyan and Batalov, 1997; Neuwald *et al.*, 1997; Park *et al.*, 1997) was NOT to use sequence-space-hopping, but to use it for reducing false positives in large-scale sequence analysis. Here, I simply applied this concept was applied to the large data set explored, and investigated its usefulness in dependence on various parameters.

More-similar-than-identical rule

A simple rule-of-thumb was explored: accept hits only if the level of sequence similarity was higher than the level of sequence identity. This rule may appear to be non-selective in that similarity would always be larger than identity; however, for the given definition of similarity (using the McLachlan metric), this was not the case.

Results

Number of false positives exploded in twilight zone

In contrast to 1990, when Sander and Schneider (1991) compiled their data, now protein pairs of dissimilar structure were detected above the 30% cut-off (Figure 2A). And these were not exceptions: at a level of 32% (HSSP-curve + 7%, i.e. n = 7 in eqn 1), the number of false positives already equalled that of homologues. For the original HSSP-curve the number of false positives was 20-fold higher than the number of true pairs. The transition from 20 to 30% sequence identity was highly non-linear for true, and false positives (logarithmic scales in Figure 2): the number of true pairs rose by a factor of 5, that of false pairs by a factor of 200 (Figure 2B). Thus, below the region of significant pairwise sequence identity (>34%) the population of false positives exploded. However, also the vast majority of homologues had less than 30% sequence identity.

Functional shape of original HSSP-curve adequate

The functional shape of the original HSSP-curve proved to be basically correct (Figure 3, grey line with triangles). However, the larger data set analysed here revealed several problems in detail (Figure 3B). (i) A threshold of 25% was not reasonable for an alignment length below 150–200 residues. (ii) Above an alignment length of about 100 residues, the derivative of the curve separating true and false positives should be lower than at lengths below 80. I attempted to solve these problems by defining a new curve for separating true and false positives (eqn 2; Figure 3, grey line with dotted circles). The particular functional form guaranteed an approximate saturation for long alignments. For alignments shorter than 11 residues eqn 2 yielded values above 100%. However, this was acceptable as 100% identity for fragments of 10–11 residues does *not* imply structural similarity (Cerpa *et al.*, 1996; Minor and Kim, 1996; Muñoz and Serrano, 1996). The new curve saturated around 20% for alignments over more than 250 residues.

Defining a curve for pairwise sequence similarity

Compiling sequence identity neglects the physico-chemical nature of amino acids. Any multiple sequence alignment illustrates that, for example, the feature hydrophobicity is more

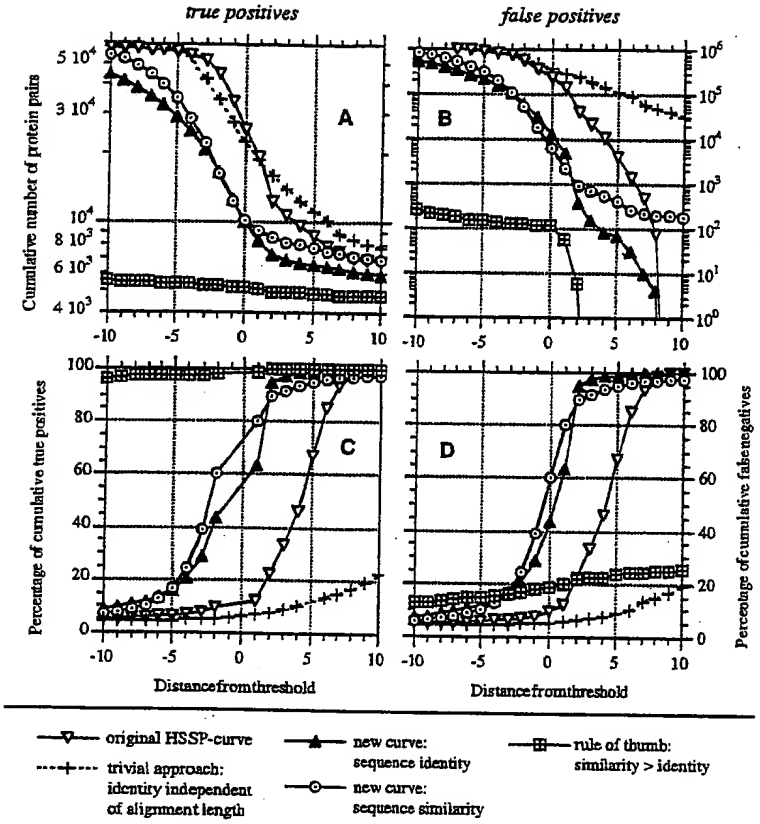


Fig. 5. Accuracy and sensitivity for detecting homologues in the twilight zone. How to choose the cut-off line for automatic database searches? The graphs A–D illustrate the pros and cons of particular choices. Given are the cumulative numbers of correctly detected homologues (true positives, **A**), and of false positives (**B**), as well as, the cumulative percentages of all correctly detected homologues (true positives, **C**), and of all homologues that were missed (false negatives, **D**) in dependence of the cut-off distance from the thresholds defined in eqn 1–3 (parameter n). Thresholds: (1) HSSP-curve (eqn 1), (2) new curve for sequence identity (eqn 2), (3) new curve for sequence similarity, (4) subset of proteins for which similarity is larger than identity (grey line in D: false negatives for this subset), (5) simple cut-off according to sequence identity disregarding alignment length (as often used in practice). Note: counts of true positives for the simple sequence identity cut-off (no alignment length) did not even fall into the interval displayed.

conserved than is the residue type. For the million protein pairs investigated here, this was reflected in a shift of the scatter plot towards lower percentages (Figure 4). In particular, for longer alignments false positives fall below 15% pairwise sequence similarity. This prompted the introduction of a threshold specifically for sequence similarity (eqn 3 in Methods; Figure 4, grey line with dotted circles). The curve surpassed 100% for alignments shorter than 12 residues and saturated at about 10% for alignments over more than 500 residues.

Better detection of homologues in twilight zone by new curves

The new curves for length-dependent cut-offs in sequence identity (eqn 2) and similarity (eqn 3) resulted in clearly lower false positive rates (higher accuracy) than the original HSSP-curve (Figure 5B and C). This was paid for by a lower number of true positives detected (lower coverage; Figure 5A). At the $n = 0$ (eqn 1–3), the old curve yielded about twofold more true positives, but more than 20-fold more false positives compared to the new curves for identity and similarity. Furthermore, at any level of true positives detected, the number of false positives was smaller for the new curves (eqn 2–3) than for the original HSSP-curve (eqn 1; Figure 7). When applying a cut-off according to mere sequence identity (ignoring alignment length), accuracy dropped below 10% at levels of 30% sequence identity (Figure 5C). Thus, detection accuracy rose almost 10-fold by the new curves.

Improving detection accuracy by expert rule

Experts often apply rules-of-thumb to visually distinguish true and false positives. However, many of such simple rules appeared not valid for automatic implementation. In particular, the distributions of the number and length of insertions did not, on average, differ between false and true positives (data not shown). Detection accuracy improved marginally by applying the following rules: (i) compile the distance for the similarity score n^S (eqn 3), and the identity score n^I (eqn 2), average over both $([n^S + n^I]/2)$, and accept pairs when this average is above some threshold n; (ii) take pairs whenever either identity or similarity surpassed the respective threshold (either $n^S \cup n^I > n$); (iii) take pairs if both values where above a given cut-off ($n^S \cup n^I > n$). In contrast, detection accuracy increased significantly by applying the 'more-similar-than-identical' rule: accept hits found in a database search only if percentage similarity is larger than percentage identity. This constraint resulted in >98% detection accuracy at $n = 0$ cut-off levels (eqn 2–3), while 2–4-fold less true positives were

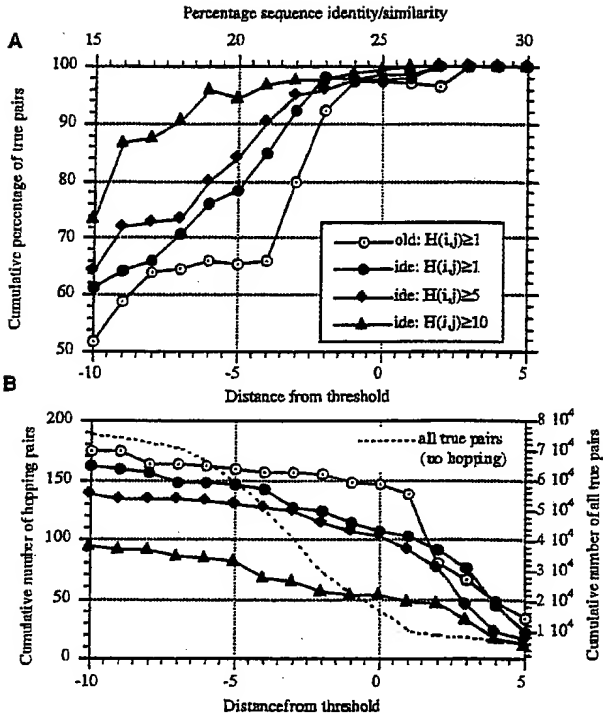


Fig. 6. Improving accuracy by sequence-space-hopping. Distances were compiled according to the old curve (eqn 1, 'old'), and to the new curve for identity (eqn 2, 'ide'). Corresponding levels of sequence identity shown on top. The cumulative percentages of true positives detected at a given cut-off distance were compiled for three different hopping strategies: hits were accepted if, at least, one (H(A,B) = 1), five (H(A,B) = 5) or 10 (H(A,B) = 10) proteins were common between two protein families (Methods). (**A**) Cumulative percentage of true positives (false positives = 100 – true); (**B**) cumulative number of true positives. The comparison of the true positives reached by intermediate sequences and all true positives (grey line in B, note: same as in Figure 2) showed that: (i) less than 1/1000 of the true positives were reached by intermediate sequences; (ii) the number of pairs reached by intermediate sequences did not explode in the twilight zone (scale on the left covers two orders of magnitude, that on the right only one). Numbers for true and false negatives would not make sense for this analysis: as we don't know all proteins, we cannot conclude that two families are unrelated only because we don't find a link between them.

found at this level (Figure 5A and C). Hence, applied as a conservative cut-off in automatic database searches, this rule proved rather powerful.

Improving detection accuracy by sequence-space-hopping

Hopping in sequence space proved successful in discarding false positives. Already the minimal constraint to accept a pair if at least one protein was common between the two sequence families yielded levels of around 80% accuracy even down to cut-off levels corresponding to 20% sequence identity (Figure 6A, compared with <20% accuracy for the normal thresholds Figure 5C). Accuracy increased further when more proteins were required to be common to both families (Figure 6A). However, sequence space hopping was possible for only relatively few protein pairs (Figure 6B). Furthermore, the improvement in accuracy was less clear using sequence-space-hopping than by applying the 'more-similar-than-identical' rule (Figure 5).

Accuracy versus coverage for BLAST and full dynamic programming

The balance between accuracy (percentage of true pairs) and coverage (percentage of all true pairs) enables choosing automatic thresholds according to a particular purpose of a database search. It also permits comparing different methods (the higher the values, the better). (i) As expected, the commonly used simple level of sequence identity (disregarding alignment length) proved, again, an extremely bad choice. (ii) Surprisingly, the fast database searching method BLAST performed relatively well in comparison to the full dynamic programming (Figure 7A). (iii) Both BLASTP version 2 and PSI-BLAST were almost as good as the full dynamic programming with the previously defined HSSP-threshold (Sander and Schneider, 1991). (iv) Best performance was achieved by the new threshold for similarity (eqn 3). (v) However, the raw alignment score performed almost as well. (vi) BLASTP (Altschul *et al.*, 1990) performed rather similarly to the more elaborate and more recent PSI-BLAST (Altschul *et al.*, 1997) (and for 'high' accuracy even slightly better, Figure 7A inset; note: given that standard parameters were chosen, this was not surprising). The corresponding thresholds were given in Figure 5B for the dynamic programming, and in Figure 7B for the PSI-BLAST probabilities.

Many false negatives at reasonable cut-off values

The number of false negatives is often of interest, i.e. the number of proteins that belong to a structure family but were not detected above a given cut-off. For the data sets used here, the cumulative percentage of false negatives was extremely high for all reasonable cut-off levels (Figure 5D). The vast majority of all pairs of proteins with similar structure populate the midnight zone below 10% sequence identity (Rost, 1997). Thus, the extremely high false negative rates proved that methods aligning two proteins merely based on the pairwise levels of sequence homology clearly fail to find the gold mine of database searches (and that older analyses that failed to describe this effect were based on biased data sets).

Thresholds for practical use

For simplicity the functions (eqn 1–3) were explicitly provided in tables (Rost, 1998). At levels of $n = 0$ (eqn 1–3) the cumulative number of true positives were (Figure 5): HSSP-curve (eqn 1), 12%; new identity curve (eqn 2), 56%; new similarity curve (eqn 3), 73%. In order to achieve levels of 99% correct hits m percentage points have to be added to the curves, where m was HSSP-curve, $m = 8$; new identity curve, $m = 5$; new similarity curve, $m = 12$. For comparison, applying the 'more-similar-than-identical' rule yielded levels above 99% down to $m = -1$.

Conclusions

Rapid transition from trivial to needle-in-haystack problem

The twilight zone of sequence pair alignments (20–35% pairwise sequence identity) was characterized by two non-linear transitions. (i) The number of homologues (true positives) rose by a factor of about eight (Figure 2A). I obtained a similar result from analysing the first four entire genomes (Rost, 1997) which indicated that this result was general, rather than database dependent. (ii) The number of false positives rose by a factor of 5000 (Figure 2B). Hence, separating true and false positives switched from a trivial task (above 35%) to the problem of finding needles in a haystack (20–30%).

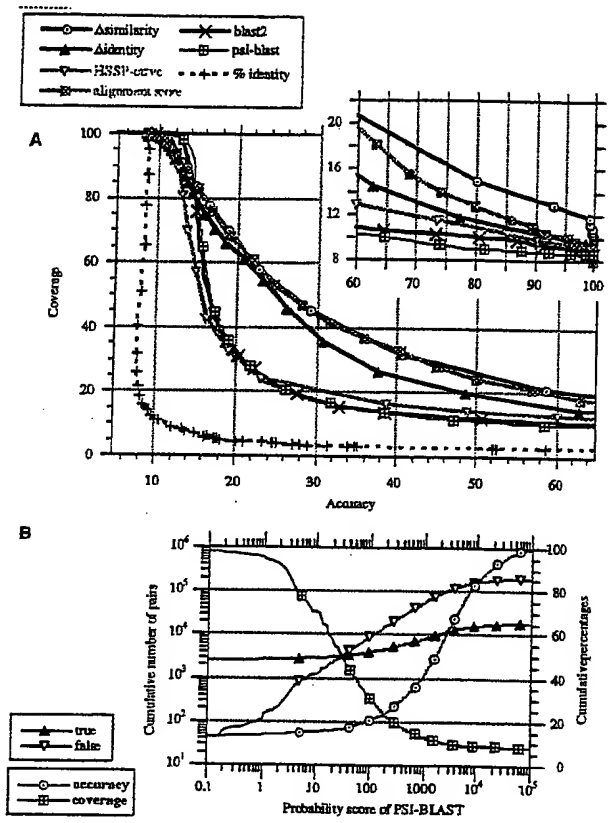


Fig. 7. Accuracy versus coverage for various methods and thresholds. Accuracy was defined as the cumulative percentage of true positives (actual true/all actual), coverage as the percentage of true positives that were detected at a given threshold (actual true/all true). (A) Thresholds and methods showed: *Δidentity*, new threshold for length-dependent sequence identity (eqn 2); *Δsimilarity*, new threshold for length-dependent sequence similarity (eqn 3); *HSSP-curve*, curve proposed by Sander and Schneider (1991; eqn 1); *%identity*, threshold given by sequence identity alone, i.e., disregarding alignment length; *alignment score*, score used for the dynamic programming optimization MaxHom; *blast2*, BLASTP version 2 (Altschul and Gish, 1996); *psi-blast*, BLASTP version 3 (Altschul *et al.*, 1997), run with standard parameters. The values for the BLAST methods were based on the probability scores reported by these algorithms. The BLAST methods did not report all pairwise alignments, thus the data set had to be reduced to the subset for which aligned pairs were reported by all three methods (MaxHom, BLASTP2, BLASTP3). Note that whereas the curves for the BLAST methods, as well as for identity and similarity are likely to hold up, in general, the curve for the alignment score is valid for the particular implementation of the dynamic programming in MaxHom, and for the particular choice of parameters (Methods). (B) Detail of the relation between the BLAST probability (here for psi-blast), and the cumulative number of true/false hits, as well as percentage accuracy and coverage.

The explosion of false positives shed light on the shape of sequence space. From 100–35% sequence identity, any residue exchange resulting in a stable structure maintains structure. From 28–35% sequence identity, most residue exchanges maintain structure. From 20–28% sequence identity, the absolute majority of residue exchanges forming stable structures populate different protein families. Is the explosion caused by features of structure space? If one generates protein sequences at random (or randomly superposes non-related proteins), the counts for most of the region above 10% sequence identity are negligible (Rost, 1997). Thus, although it is obvious that we expect to find more pairs for lower levels of sequence identity based on mere statistics, the particular transition in the twilight zone seems not to be evident. However, this analysis did not provide answers to whether or not the observed explosion may reflect structural (Chung and Subbiah, 1996) and/or functional constraints.

Poor distinction between true and false positives by sequence identity, alone

Even journals such as Cell, or EMBO provide an ample source for the following fallacy: 'these two fragments of 16 residues adopt similar structures as they have more than 10 similar residues'. Thus, one of the most important messages of this analysis might be the repetition of a point made by others (Sander and Schneider, 1991): high levels of sequence similarity or identity do *not* ascertain structural similarity (Figure 5). Instead, the levels of significant sequence identity and similarity depend on the alignment length (Figures 3 and 4), or the respective raw score of the alignment methods.

Better distinction by new curves for sequence identity and similarity

The length-dependent cut-off for significant sequence identity pioneered by Sander and Schneider (1991) needed refinement in several ways to account for the findings from a 1000-fold larger data set: (i) shift towards higher values for shorter alignments; (ii) saturation for alignments longer than 150 residues; (iii) definition of new curve for levels of sequence similarity. These tasks were solved by introducing threshold curves for significant sequence identity (eqn 2), and for significant sequence similarity (eqn 3). The precise definition of the two thresholds was entirely empirical. However, the essential functional dependency of the curves was kept similar to what would be expected from pure statistical considerations. Although not true for all problems (Nielsen *et al.*, 1996), on average, sequence similarity was marginally more successful than identity in distinguishing true and false positives. The new curves improved accuracy at a given coverage (Figure 5 and 7). Additionally, this analysis supplied detailed levels for expected accuracy and coverage for the curves defined, as well as for standard BLAST searches (Figures 5 and 7). Such estimates may have implications for automatic database searches. They also shed light on the comparison between sequence alignments and threading techniques that both only make use of pair comparisons (rather than using family specific profiles): already at levels of 25% sequence identity, pair alignments detect only 10–30% true positives. This is below the level of what threading techniques achieve in the interval 0–25% sequence identity (Sippl, 1995; Fischer and Eisenberg, 1996; Russell *et al.*, 1996; Rost *et al.*, 1997).

Improved accuracy by 'more-similar-than-identical' rule and sequence space hopping

The number of false positives was significantly reduced by two techniques (only the first of which was novel to this work). (i) The 'more-similar-than-identical' rule: 95% of all pairs for which percentage similarity was larger than percentage identity had similar structures. Thus, this constraint clearly improved detection accuracy. The cost was low coverage: for only 10% of the structurally similar pairs the percentage similarity was larger than percentage identity. This might be explained by the fact that half of the protein, on average, embedded in loop regions, may tolerate residue exchanges that do not conserve physico-chemical properties (and thus decrease

the overall average more than the few to-be-conserved-regions increase it). (ii) The usage of 'multi-links' (Abagyan and Batalov, 1997), 'intermediate sequences' (Park *et al.*, 1997), 'transitivity' (Neuwald *et al.*, 1997), or 'sequence space hopping': most protein pairs that contained a similar subset of identical proteins in their respective sequence families were found to have similar structures even at low levels of sequence homology. Obviously, the validity of transitivity (detection accuracy) between protein families (Figure 1) depended on the distance between the families (Figure 6). Interestingly, the improvement of accuracy hardly depended on the number of proteins required to be common to two families. This suggested that although the vast majority of protein pairs with 25% sequence identity had dissimilar structures, the 'islands' populated by structure families were well separated. Unfortunately, for the data set explored here, the yield of this analysis was found to be very low: on average only one in 1000 pairs was reached via intermediate sequences (Figure 6). Furthermore, sequence-space-hopping resulted in clearly lower coverage/accuracy ratios than did the application of the 'more-similar-than-identical' rule (Figures 5 and 6).

Beginning of the 90's: over-estimation of sequence alignment methods

Until 1996, very few people had taken up the laborious task of objective large-scale analyses of protein sequence comparisons. Partially, because automatic structure comparison methods are fairly recent. The few earlier workers (Sander and Schneider, 1991; Vogt *et al.*, 1995; Gotoh, 1996) based their work on data sets of about 1000 pairs of protein structure alignments. Gotoh (1996) and Vogt *et al.* (1995) used the same set (Pascarella and Argos, 1992) for testing different alignment methods, and a variety of substitution metrices. They focused on monitoring the detailed accuracy in terms of number of residues correctly aligned. Due to the small data set Vogt *et al.* (1995) found about 98% true positives at 30% sequence identity (ignoring alignment length), and 50% true positives at 20% sequence identity. For the 1000-fold larger data set used here the corresponding values were quite different (ignoring alignment length): 11% true positives at 30% sequence identity, and 5% true positives at 20% identity. However, even the more conservative analysis introducing the importance of alignment length for levels of significant sequence identity (Sander and Schneider, 1991) still over-estimated the possible levels of sequence identity between proteins of dissimilar structure.

End of the 90's: database searches do not reach the gold mine, yet

The thresholds for sequence identity and similarity defined here, as well as those established by others (Abagyan and Batalov, 1997; Brenner *et al.*, 1998) complemented the levels for 'significance' provided by BLAST (Altschul and Gish, 1996), FASTA (Pearson, 1996) or other statistical analyses (Bryant and Altschul, 1995) by addressing the question 'how significant is the significance of the respective alignment method?'. Based on quite different data sets the principal messages were similar: (i) most proteins of similar structure were not found by pairwise sequence comparisons at reasonable cut-off thresholds; (ii) raw scores from dynamic programming methods were comparable to the original length-dependent cut-off thresholds for sequence identity (Sander and Schneider, 1991); (iii) dynamic programming was only slightly superior to BLAST searches (Altschul and Gish, 1996; Altschul *et al.*, 1997). However, in detail the numbers differed between the recent analyses. Obviously, the absolute values depended crucially on the particular choice of the data set. Abagyan and Batalov (1997) analysed various substitution metrices on a data set comparable to the one used in this analysis. They concluded that raw alignment scores provide better separations between true and false positives than do length-dependent cut-offs for sequence identity and similarity. The difference between their result, and the one shown here may result from the fact that Abagyan and Batalov (1997) used the optimal choice of all parameters for comparing the raw alignment score to sequence identity and similarity. Brenner and co-workers have analysed the accuracy and coverage for various statistical scores (Brenner *et al.*, 1998). They used a completely different data set than I did. An approximate comparison of the two analyses was possible by the reference point of simple identity (ignoring alignment length). It seems that the performance for the best separation method they find (new FASTA) was comparable to the improved, simple thresholds defined here (eqn 2–3). Here, the BLAST probability was found to be a relatively good way to separate true and false positives (Figure 7A): it was only slightly inferior to the raw dynamic programming alignment score, results for which hold up exclusively for the particular choice of parameters and the particular alignment algorithm used.

Thresholds in practice

The advantages of the length-dependent levels of identity and similarity (eqn 2–3) over other thresholds (Abagyan and Batalov, 1997; Alexandrov and Soloveyev, 1998) was that these thresholds, in principle, are applicable to any alignment, and may relate more explicitly to structure. Identity and similarity can be compiled easily without having to re-do the entire database search. In practice, this does not always hold up: (i) different parameters (e.g. the way in which gaps are treated) may result in different alignments; and (ii) the similarity values compiled hold for the choice of a particular metric (here McLachlan). Additionally, the thresholds introduced here provide independent evidence for the separation, and permitted the application of the successful 'more-similar-than-identical' rule.

Will the analysis hold up for the next 500 structures?

The results given here based on the largest possible data set for which structural alignments provided a well-defined distinction between true and false. One conclusion was that seven years ago (Sander and Schneider, 1991) the database was too small to capture the details. Will this also be true in 2005? Answers have to remain speculative. (i) Although the database used in 1990 was 1000-fold smaller than the one used here, some principle findings were verified. (ii) Assuming that there are only 1000 folds in nature (Chothia, 1992), and that these correspond to about 10 000 families, then even the full catalogue of all protein sequences would yield a data set essentially only 30 times larger than the one used here (note: the data set used corresponded to about 300 different folds aligned against about 1000 families).

Rather more accurate, or more sensitive?

An accurate and sensitive distinction between true and false positives is important for automatic database searches. The new curves introduced here (eqn 2–3) proved slightly more sensitive (higher coverage) and more accurate than the previously proposed curve (Sander and Schneider, 1991). The

accuracy increased significantly by applying the 'more-similar-than-identical' rule, and by sequence space hopping. However, accuracy was gained at the expense of coverage. Which is more important? Clearly, the evolutionary information contained in multiple alignments is the single most important contribution to improving protein structure prediction in the 90's (Rost and Sander, 1996; Rost and O'Donoghue, 1997). Is the gain by increased diversity more important than the loss of accuracy when using alignments for structure prediction? The answer depends on the particular prediction goal. For example, for secondary structure prediction diversity is more important than accuracy (cut-off at 25% versus that at 30%), whereas for the prediction of solvent accessibility the opposite is true (unpublished). Furthermore, as databases grow coverage may be less important than accuracy. Irrespective of individual preferences, the sharper the knife cutting between true and false positives, the better. This analysis has sharpened the knife a little, and added new optional tools to it.

Acknowledgements

Particular thanks to Richard Friedman (Columbia, New York) for his extremely helpful comments on the manuscript. Furthermore, thanks to Reinhard Schneider (LION, Heidelberg) for discussions; to Jong Park (MRC, Cambridge) and Tim Hubbard (Sanger, Hinxton, UK) for providing their manuscript prior to publication, and for the two anonymous referees for their helpful comments. Financial support was provided by Matti Sarraste (EMBL, Heidelberg), Gerrit Vriend (EMBL, Heidelberg), Chris Sander (Millenium, Boston) and Friedrich von Bohlen (LION, Heidelberg). Last, but not least, thanks to all those who deposit experimental information about protein structures and sequences in public databases, and to those maintaining these sources of knowledge.

References

Abagyan,R.A. and Batalov,S. (1997) *J. Mol. Biol.*, **273**, 355–368.
Alexandrov,N.N. and Soloveyev,V.V. (1998) In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds) *HICCS' 98: Pacific Symposium on Biocomputing' 98*. Maui, Hawaii, USA, pp. 463–472.
Alexandrov,N.N., Takahashi,K. and Go,N. (1992) *J. Mol. Biol.*, **225**, 5–9.
Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A.G. (1996) *Methods Enzymol.*, **266**, 635–643.
Bryant,S.H. and Altschul,S.F. (1995) *Curr. Opin. Struct. Biol.*, **5**, 236–244.
Casari,G., Sander,C. and Valencia,A. (1995) *Nature Struct. Biol.*, **2**, 171–178.
Cerpa,R., Cohen,F.E. and Kuntz,I.D. (1996) *Folding Des.*, **1**, 91–101.
Chothia,C. (1992) *Nature*, **357**, 543–544.
Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.
Chung,S.Y. and Subbiah,S. (1996) *Structure*, **4**, 1123–1127.
Cohen,B.I., Presnell,S.R. and Cohen,F.E. (1993) *Protein Sci.*, **2**, 2134–2145.
Crippen,G.M. and Maiorov,V.N. (1995) *J. Mol. Biol.*, **252**, 144–151.
Doolittle,R.F. (1979) In Neurath,H. and Hill,R.L. (eds) *Protein Evolution*. Academic Press, New York, pp. 1–118.
Doolittle,R.F. (1981) *Science*, **214**, 149–159.
Doolittle,R.F. (1986) *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA, USA.
Doolittle,R.F. (1994) *TIBS*, **19**, 15–18.
Fischer,D. and Eisenberg,D. (1996) *Protein Sci.*, **5**, 947–955.
Fischer,D., Rice,D.W., Bowie,J.U. and Eisenberg,D. (1996) *FASEB J.*, **10**, 126–136.
Gerstein,M. and Levitt,M. (1996) In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R.F. (eds) *Fourth International Conference on Intelligent Systems for Molecular Biology*, AAAI, St Louis, MO, USA, pp. 59–67.
Gotoh,O. (1996) *J. Mol. Biol.*, **264**, 823–838.
Gribskov,M., McLachlan,M. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–5358.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Holm,L., Ouzounis,C., Sander,C., Tuparev,G. and Vriend,G. (1993) *Protein Sci.*, **1**, 1691–1698.
Holm,L. and Sander,C. (1996) *Nucleic Acids Res.*, **25**, 231–234.
Holm,L. and Sander,C. (1997) *Proteins*, **28**, 72–82.
Holmes,K.C., Sander,C. and Valencia,A. (1993) *TICB*, **3**, 53–59.
Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.
Kabsch,W. and Sander,C. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.
Lemer,C.M.-R., Rooman,M.J. and Wodak,S.J. (1995) *Proteins*, **23**, 337–355.
Luo,Y., Lai,L., Xu,X. and Tang,Y. (1993) *Protein Engng*, **6**, 373–376.
Maiorov,V.N. and Crippen,G.M. (1995) *Proteins*, **22**, 273–283.
Minor,D.L.J. and Kim,P.S. (1996) *Nature*, **380**, 730–734.
Muñoz,V. and Serrano,L. (1996) *Folding Des.*, **1**, R71–R77.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) *Nucleic Acids Res.*, **25**, 1665–1677.
Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) *Proteins*, **24**, 165–177.
Orengo,C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 429–440.
Orengo,C.A. and Taylor,W.R. (1996) *Meth. Enzymol.*, **266**, 617–635.
Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.
Orengo,C.A., Michie,A.D., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structures*, **5**, 1093–1108.
Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) *J. Mol. Biol.*, **273**, 349–354.
Pascarella,S. and Argos,P. (1992) *Protein Engng*, **5**, 121–137.
Pearson,W.R. (1996) *Methods Enzymol.*, **266**, 227–258.
Rost,B. (1996) *Methods Enzymol.*, **266**, 525–539.
Rost,B. (1997) *Folding Des.*, **2**, S19–S24.
Rost,B. and O'Donoghue,S.I. (1997) *CABIOS*, **13**, 345–356.
Rost,B. and Sander,C. (1996) *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 113–136.
Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.
Rost,B. (1997) WWW document (http://www.embl-heidelberg.de/predictprotein). EMBL.
Rost,B. (1998) WWW document (http://www.embl-heidelberg.de/~rost/Papers/Dfig/98twilight/app.html). EMBL.
Russell,R.B., Copley,R.R. and Barton,G.J. (1996) *J. Mol. Biol.*, **259**, 349–365.
Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
Schneider,R. (1994) PhD, University of Heidelberg.
Schneider,R., de Daruvar,A. and Sander,C. (1997) *Nucleic Acids Res.*, **25**, 226–230.
Sippl,M.J. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.
Sippl,M.J. and Floeckner,H. (1996) *Structure*, **4**, 15–19.
Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
Valencia,A., Kjeldgaard,M., Pai,E.F. and Sander,C. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 5443–5447.
Vogt,G., Etzold,T. and Argos,P. (1995) *J. Mol. Biol.*, **249**, 816–831.
Wodak,S.J. and Rooman,M.J. (1993) *Curr. Opin. Struct. Biol.*, **3**, 247–259.
Zu-Kang,F. and Sippl,M.J. (1996) *Folding Des.*, **1**, 123–132.
Zuckerkandl,E. (1976) *J. Mol. Evol.*, **7**, 269–311.
Zuckerkandl,E. and Pauling,L. (1965) In Bryson,V. and Vogel,H.J. (eds) *Evolutionary Divergence and Convergence in Proteins*. Academic Press, New York; London, pp. 97–166.

Received March 23, 1998; revised October 23, 1998; accepted October 27, 1998

EXHIBIT 7

CLUSTAL W (1.8) multiple sequence alignment

```
gi|4406210|gb|AAD19901.1|     ----------------------------------------------------
gi|2661771|emb|CAA74140.1|    ----------MAEFEERPRLSIGEEEAPPYPLEKLTGGRRRTRAQIHALH
gi|4416322|gb|AAD20314.1|     ----------------------------------------------MSE
gi|12642370|gb|AAK00231.1|AF22 ----------------------------------MNEITTSEQLDLQ
9306aa.txt                    WKIDNRELSLPTFCGCAQPSTIHFLYFNKIQMAEQRRNTRSTAKSKVQPV

gi|4406210|gb|AAD19901.1|     -MEGPIPPHSLEAEQSVLGSILLDSDVMDEVEGLLPSPEAFYAEAHRKIY
gi|2661771|emb|CAA74140.1|    QQAGRVPPQAVELEQAVLGAMLIEPEAIPRALEILT-PEAFYDGRHQRIF
gi|4416322|gb|AAD20314.1|     LFSERIPPQSIEAEQAVLGAVFLDPAALVPASEILI-PEDFYRAAHQKIF
gi|12642370|gb|AAK00231.1|AF22 TAALKVPPHSIEAEQAVLGGLMLDNNAWERVLDQVS-DGDFYRHDHRLIF
9306aa.txt                    NDYGRIQPQAPELEEAVLGALMIEKDAYSLVSEILR-PESFYEHRHQLIY
                                   : *:: * *::***.:::: .  .  :    **   *: *:

gi|4406210|gb|AAD19901.1|     AAMQALRSQGRPVDLVTLSEELSRRGQLEEVGGTAYLLQLSEATPTAAYA
gi|2661771|emb|CAA74140.1|    RAIVRLFEQNRGVDLLTVTEELRRTGELEQAGDTIYLSELTTRVASAANV
gi|4416322|gb|AAD20314.1|     HAMLRVADRGEPVDLVTVTAELAASEQLEEIGGVSYLSELADAVPTAANV
gi|12642370|gb|AAK00231.1|AF22 RAVHKLADANQPFDVVTLHEQLDKEGLSSQVGGLAYLAELAKNTPSVANI
9306aa.txt                    AAITDLAVNQKPVDILTVKEQLSKRGELEEVGGPFYITQLSSKVASSAHI
                               *:  :    . .*::*:  :*      .: *.  *: :*:  ..: *

gi|4406210|gb|AAD19901.1|     EHYARIVAEKWTLRRLIQAAGEAMRLAYEEAG-SLDEILDTAGKKILEVA
gi|2661771|emb|CAA74140.1|    EYHARIIAEKALLRRMIEVMTLLVGRAYDPAA-DAFELLDEVEAEIFRLS
gi|4416322|gb|AAD20314.1|     EYYARIVEEKSVLRRLIRTATSIAQDGYTRED-EIDVLLDEADRKIMEVS
gi|12642370|gb|AAK00231.1|AF22 KAYAAIIRERATLRQLISISTDIADNAFNPQGRNAAEILDDAERQIFQIA
9306aa.txt                    EYHARIIAQKYLARELITFTSNIQSKAFDETL-DVDDLMQEAEGKLFEIS
                              : :* *: ::   *.:*          .:     .   ::: .  :::.::

gi|4406210|gb|AAD19901.1|     LTKTDTEARP-MRELVHETFEHIEALFQNKGEVAGVRTGFKELDQLIGTL
gi|2661771|emb|CAA74140.1|    DVHLRKAARS-MNEVVKETLERLEAIHGRPGGITGVPSGFHQLDALTGGW
gi|4416322|gb|AAD20314.1|     QRKHSGAFKN-IKDILVQTYDNIEMLHNRDGEITGIPTGFTELDRMTSGF
gi|12642370|gb|AAK00231.1|AF22 EARPKTGGPVGVNELLTMAIDRIDTLFNSDSDITGISTGYTDLDEKTSGL
9306aa.txt                    QRNMKKDYTQ-INPIIAEAYEQIQKAAARTDGLSGLESGYTKLDKMTSGW
                                .        :. ::  : :.::       . ::*: :*: .**   .

gi|4406210|gb|AAD19901.1|     GPGSLNIIAARPAMGKTAFALTIAQNAALK--EGVGVGIYSLEMPAAQLT
gi|2661771|emb|CAA74140.1|    QRGDLIIIAARPSMGKTAFALSCARNAALHPHYGTGVAIFSLEMGAEQLA
gi|4416322|gb|AAD20314.1|     QRSDLIIVAARPSVGKTAFALNIAQNVATK--TNENVAIFSLEMSAQQLV
gi|12642370|gb|AAK00231.1|AF22 QAADLIIVAGRPSMGKTTFAMNLVENAVLR--TDKAVLVFSLEMPGESLI
9306aa.txt                    QKSDLIIIAARPAMGKTAFVLSMAKNIAVN--FRNPVALFSLEMSNVQLV
                                ..* *:*.**::***:*.:. ..* . .       * ::****   .*

gi|4406210|gb|AAD19901.1|     LRMMCSEARIDMNRVRLGQLTDRDFSRLVDVASRLSEAPIYIDDTPDLTL
gi|2661771|emb|CAA74140.1|    QRLLTAEARVDAQAARTGRLRDEDWRKLARAAGRLSDAPIFIDDTPSLGV
gi|4416322|gb|AAD20314.1|     MRMLCAEGNINAQNLRTGKLTPEDWGKLTMAMGSLSNAGIYIDDTPSIRV
gi|12642370|gb|AAK00231.1|AF22 MRMLSSLGRIDQTKVRSGQLDDDDWPRLTSAVNLLNDRKLFIDDTAGISP
9306aa.txt                    NRLISNVCEIPSEKIKSGQLAAYEWQQLDYKLKDLLDAPLYVDDTPSLSV
                               *::     .:      : *:*   :: :*       * :  :::***..:

gi|4406210|gb|AAD19901.1|     MEVRARARRLVSQN-QVGLIIIDYLQLMSGPGSGKSGENRQQEIAAISRG
gi|2661771|emb|CAA74140.1|    LELRAKCRRLKAEH-DIGLVIVDYLQLMQASHMPRN-ANREQEIAQISRS
gi|4416322|gb|AAD20314.1|     SDIRAKCRRLKQES-GLGMIVIDYLQLIQGSGRSKE--NRQQEVSEISRS
gi|12642370|gb|AAK00231.1|AF22 SEMRARTRRLAREHGEIAMIMVDYLQLMQIPGSAGD--NRTNEISEISRS
9306aa.txt                    FELRTKARRLVREH-GVRIIIIDYLQLMNASGMAFG--SRQEEVSTISRS
                               ::*:: ***  :   : ::::*****:. .        .* :*:: ***.
```

```
gi|4406210|gb|AAD19901.1|     LKALARELGIPIIALSQLSRAVEARP---NKRPMLSDLRES---------
gi|2661771|emb|CAA74140.1|    LKALAKELNVPVVALSQLSRAVETRGG--DKRPQLSDLRESGCLAGDTLI
gi|4416322|gb|AAD20314.1|     LKALARELEVPVIALSQLSRSVEQRQ---DKRPMMSDIRES---------
gi|12642370|gb|AAK00231.1|AF22 LKALAKEFNCPVIALSQLNRSLEQRP---NKRPVNSDLRES---------
9306aa.txt                    LKGLAKELNIPIIALSQLNRGVESREGLEGKRPQLSDLRES---------
                              **.**:*:  *::*****.*.:* *    .***  **:***

gi|4406210|gb|AAD19901.1|     ------G-----------------------------------------------
gi|2661771|emb|CAA74140.1|    TLADGRRVPIRELVSQQNFSVWALNPQTYRLERARVSRAFCTGIKPVYRL
gi|4416322|gb|AAD20314.1|     ------G-----------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 ------G-----------------------------------------------
9306aa.txt                    ------G-----------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    TTRLGRSIRATANHRFLTPQGWKRVDELQPGDYLALPRRIPTASTPTLTE
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    AELALLGHLIGDGCTLPHHVIQYTSRDADLATLVAHLATKVFGSKVTPQI
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    RKELRWYQVYLRAARPLAPGKRNPISDWLRDLGIFGLRSYEKKVPALLFC
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    QTSEAIATFLRHLWATDGCIQMRRGKKPYPAVYYATSSYQLARDVQSLLL
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    RLGINARLKTVAQGEKGRVQYHVKVSGREDLLRFVEKIGAVGARQRAALA
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    SVYDYLSVRTGNPNRDIIPVALWYELVREAMYQRGISHRQLHANLGMAYG
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------

gi|4406210|gb|AAD19901.1|     -----------------------------------------------------
gi|2661771|emb|CAA74140.1|    GMTLFRQNLSRARALRLAEAAACPELRQLAQSDVYWDPIVSIEPDGVEEV
gi|4416322|gb|AAD20314.1|     -----------------------------------------------------
gi|12642370|gb|AAK00231.1|AF22 -----------------------------------------------------
9306aa.txt                    -----------------------------------------------------
```

```
gi|4406210|gb|AAD19901.1|          --------------------SIEQDADLVMFIYRDEYYNPHSEKAG----
gi|2661771|emb|CAA74140.1|         FDLTVPGPHNFVANDIIAHNSIEQDADVVLFIYRPERYGITVDENGNPTE
gi|4416322|gb|AAD20314.1|          --------------------SIEQDADIVAFLYRDDYYNKDSENKN----
gi|12642370|gb|AAK00231.1|AF22     --------------------AIEQDADVIMFVYRDEVYHPETEHKG----
9306aa.txt                         --------------------AIEQDADMVCFIHRPEYYKIFQDDKGNDLR
                                                       :******:: *::* : *    :. .

gi|4406210|gb|AAD19901.1|          -IAEIIVGKQRNGPTGTVELQFHASHVRFND---------------LARD
gi|2661771|emb|CAA74140.1|         GIAEIIIGKQRNGPTGTVRLAFINQYARFEN---------------LTMY
gi|4416322|gb|AAD20314.1|          -IIEIIIAKQRNGPVGTVQLAFIKEYNKFVN---------------LERR
gi|12642370|gb|AAK00231.1|AF22     -VAEIIIGKQRNGPIGFVRLAFIGKYTRFEN---------------LAPG
9306aa.txt                         GMAEIIIAKHRNGAVGDVLLRFKGEYTRFQNPDDDMVIPLPDAGAMLGSR
                                    : ***:.*:***. * * * *  .: :* :                *

gi|4406210|gb|AAD19901.1|          A------------------------------
gi|2661771|emb|CAA74140.1|         QPEPGTPLPETPDETILPSGPPDEAPF-----
gi|4416322|gb|AAD20314.1|          FDEAQIPPGA----------------------
gi|12642370|gb|AAK00231.1|AF22     MYNFDDDE---------------------------
9306aa.txt                         MNNTGTVPPPPAEFAPQNSNPFGGENDGPLPF
```